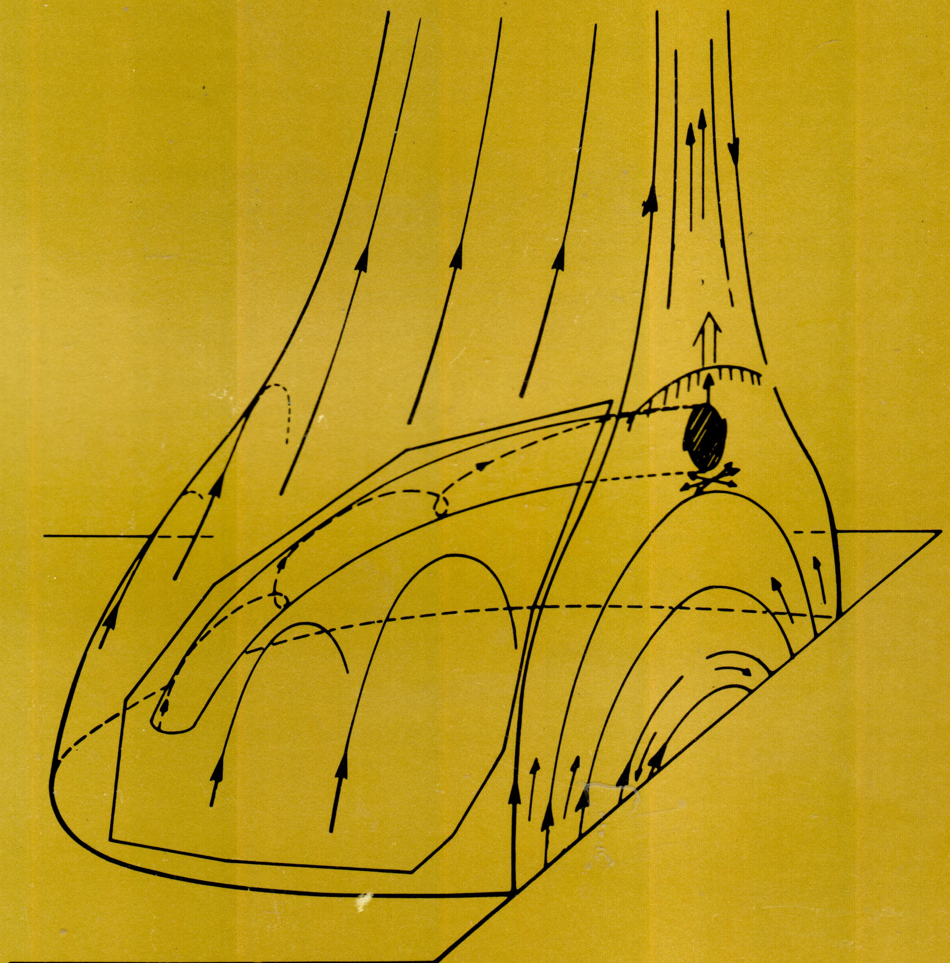
Nederlands Tijdschrift voor Natuurkunde

NNV

Uitgave van de
Nederlandse Natuurkundige Vereniging

A 46 (1 1980)

Inhoud



Nederlands Tijdschrift voor Natuurkunde

Uitgave van de Nederlandse Natuurkundige Vereniging NNV

Jaargang A46 (1), maart 1980

Produktie en advertenties:
Bureau Ellens BV,
Duinweg 9 (2585 JT), Postbus 84323
(2508 AH), Den Haag, telefoon 070-501205*

Druk:
Samsom-Sijthoff Grafische Bedrijven
Alphen aan den Rijn

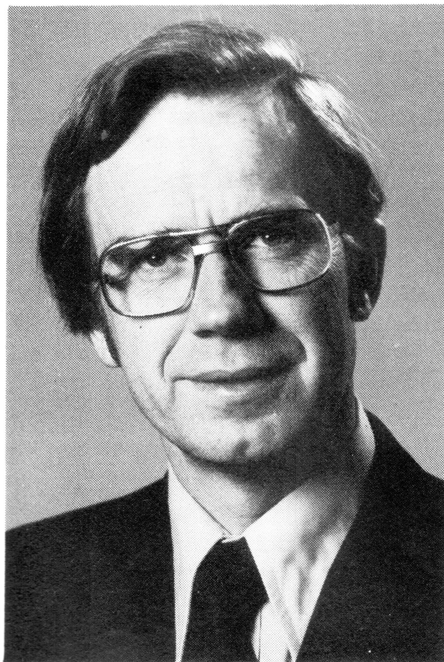
Abonnementen:
De Nederlandse Natuurkundige Vereniging, Princetonplein 5, Postbus 80.000, 3508 TA Utrecht, telefoon 030-532329, postgiro 263079. De administratie behandelt alle zaken betreffende het lidmaatschap van de NNV, abonnementen op het NTvN, evenals alle betalingen.
Het Nederlands Tijdschrift voor Natuurkunde-A verschijnt viermaal per jaar en is hoofdzakelijk bestemd voor korte overzichtsartikelen, die van belang kunnen zijn voor een in de fysica geïnteresseerde lezerskring.
Actuele informatie op natuurkundig en hiermee samenhangend maatschappelijk gebied, verenigingsnieuws en verdere mededelingen aan de Nederlandse natuurkundigen worden gegeven in het veertiendaagse tijdschrift NTvN-B.
Beide tijdschriften worden uitgegeven door de Nederlandse Natuurkundige Vereniging en toegezonden aan haar leden.
De contributie van de NNV bedraagt bij gewoon lidmaatschap f 79,–; voor leden van de KNCV, de NAC, NVON, de Belg. Nat. Ver. of het KNGMG f 67,–; voor studenten en militairen f 26,–; voor afgestudeerden gedurende 2 jaar na het afstuderen f 53,–.
Leden van de KNCV, de NAC, het Wiskundig Genootschap, NVON of het NERG, kunnen zich abonneren op het NTvN tegen de gereduceerde prijs van f 47,50. Opgave voor gereduceerde abonnementen bij het secretariaat van de betreffende vereniging.
Betaling geschiedt aan de NNV.
Losse abonnementen op het NTvN kunnen worden genomen bij de administratie van de NNV.
Abonnementsprijs f 63,– (buitenland f 78,–).
Losse nummers f 5,50.

Figuur voorpagina: Schematische voorstelling van de gebeurtenissen tijdens een zonnevlam.

De eerste aflevering van het NTvN-A in 1980, dat thans voor u ligt, begint met een korte bijdrage van het bestuur van de NNV, waarin een recente redactie-wijziging wordt gemeld. Daarin worden ook de verdiensten van de aftredende redactieleden voor dit tijdschrift gememoreerd. Zodoende kan ik mij hier beperken tot de opmerking dat m.i. de oude redactie er volledig in is geslaagd om de enkele jaren geleden nieuw geformuleerde doelstelling van het NTvN-A – het vele interessante speurwerk dat in Nederland of elders gebeurt op natuurkundig of verwant gebied in korte overzichtsartikelen onder de aandacht van de leden van de NNV te brengen – te realiseren. Het was verheugend dat zoveel onderzoekers bereid zijn geweest voor het NTvN-A over hun werk te schrijven. Ieder die wel eens voor de taak heeft gestaan om de essentie van een verricht onderzoek duidelijk weer te geven voor een lezerskring zoals die van dit tijdschrift met een wijd uiteenlopende fysische belangstelling, weet dat dit geen eenvoudige opgave is. Het vereist misschien wel meer inspanning en inzicht dan het schrijven van een artikel voor een vaktijdschrift. De redactie hoopt ook in de toekomst op dezelfde enthousiaste medewerking te kunnen rekenen.
Dit nummer bevat een zestal artikelen op het gebied van de halfgeleiderfysica en -technologie, een in eerste instantie doelgericht onderzoeksgebied met vele interessante fundamentele en praktische problemen, zoals u bij lezing van de artikelen duidelijk zal worden. Deze artikelen zijn tot stand gekomen met de medewerking van de sectie Halfgeleiderfysica en -elektronica van de NNV.
De overige artikelen zijn gewijd aan een viertal onderwerpen uit de astrofysica en geven een duidelijke illustratie van hoe boeiend dit gebied van onderzoek is.

A. van der Woude





De redactie van het NTvN-A heeft enkele veranderingen ondergaan. Dr. J. Polman heeft afscheid genomen als redactievoorzitter. Hij wordt als zodanig opgevolgd door dr. A. van der Woude, KVI, Rijksuniversiteit Groningen. Ook dr. P.W.M. Glaudemans heeft zijn taak als lid van de redactie neergelegd. Als nieuwe leden treden toe dr. A. Broese van Groenou, Philips Natuurkundig Laboratorium, Eindhoven en dr. H.M. Rietveld, ECN, Petten, waardoor de redactie dus met één lid is uitgebreid. Bij deze gelegenheid past een woord van oprechte dank van het NNV-bestuur aan Glaudemans en Polman.

Piet Glaudemans, 48, hoogleraar in de kernfysica aan de Rijksuniversiteit te Utrecht, trad in oktober 1970 toe tot de toenmalige redactieraad. Door problemen van verschillende aard werd het in het midden van de zeventiger jaren steeds moeilijker bij de toenmalige opzet van het tijdschrift de gewenste kwaliteit te handhaven. Tegelijkertijd begon de groei van het ledental van de vereniging te stagneren, ja zelfs trad er een lichte daling op (zie NNV-jaarboek 1978/79, p. 8). Een samenhang met de moeilijkheden van het tijdschrift werd vermoed. Glaudemans, in 1974 secretaris van de NNV geworden, heeft zich toen met hart en ziel ingespannen voor een verbetering in de situatie, hetgeen in januari 1977 heeft geleid tot de ons nu vertrouwde splitsing van het tijdschrift in de delen A en B. Hij werd lid van de redactie van het NTvN-A. Het goede onthaal van dit blad bij de leden en de snelle stijging van het ledental in de jaren 1977 en 1978 hebben hem terecht veel voldoening gegeven.

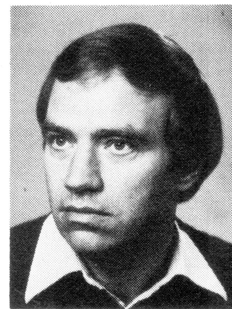
Jan Polman, 44, verbonden aan het Philips Natuurkundig Laboratorium te Eindhoven, werd in januari 1972 lid van de redactieraad en twee jaar later algemeen redacteur. Ook hij was nauw betrokken bij het overleg over de reorganisatie. Zijn brede belangstelling voor alle aspecten van de natuurkunde maakte hem bij uitstek geschikt voor het redactievoorzitterschap. In deze functie, die hij in januari 1977 aanvaardde, is hij erin geslaagd een mooi evenwicht te vinden tussen wetenschappelijke strengheid en popularisering, tussen aandacht voor fundamenteel onderzoek en voor technische toepassingen, terwijl daarnaast ook het onderwijs en de problemen van natuurkunde en samenleving niet werden vergeten.

Glaudemans en Polman hebben in de afgelopen jaren in belangrijke mate het gezicht van het NTvN bepaald. Hun werk wordt nu overgenomen door anderen. Het bestuur prijst zich zeer gelukkig met de nieuwe bemanning en wenst haar veel succes. Het doet tevens een beroep op de leden de taak van de redactie te verlichten door een enthousiaste en spontane medewerking.

Q.H.F. Vrehen
voorzitter redactie-adviesraad

Lithografie ten behoeve van Very Large Scale Integration

R.P. Kramer



Drs. R.P. Kramer (36) studeerde experimentele natuurkunde aan de Gemeente Universiteit van Amsterdam. Hij is sinds 1970 werkzaam als wetenschappelijk medewerker van het Philips Natuurkundig Laboratorium te Eindhoven met als vakgebied halfgeleiderfysica en -technologie. De laatste jaren houdt hij zich vooral bezig met problemen die samenhangen met het verkleinen van MOS-schakelingen, in het bijzonder met de lithografie hiervan.

Het toenemen van de complexiteit van geïntegreerde schakelingen gaat enerzijds gepaard met het groter worden van het chip-oppervlak, anderzijds met het kleiner worden van de minimale afmetingen op de chip. De eisen die worden gesteld aan de patroondefinitie worden daardoor strenger. In dit artikel wordt een overzicht gegeven van verschillende lithografische methoden die ons in deze ter beschikking staan; speciale aandacht zal worden besteed aan de licht-optische repeteerprojector en de elektronenoptische beeldprojector.

De vervaardiging van een transistor of geïntegreerde schakeling bestaat uit de opeenvolging van een aantal lithografische stappen, afgewisseld door een combinatie van etsen, oxyderen, implanteren, diffunderen en deponeren van diëlektrische of geleidende laagjes. Onder lithografie verstaan we de verzameling processtappen die ertoe leidt dat een gewenst patroon op het oppervlak van een substraat wordt overgebracht. In de IC(Integrated Circuits)-technologie gebeurde dat tot voor kort vrijwel uitsluitend door een masker, met daarop het gewenste patroon gerepeteerd aanwezig, in nauw contact te brengen met een plak silicium waarop een laagje fotogevoelige lak was aangebracht en deze plak-maskercombinatie met ultraviolet licht te belichten (het zogenaamde *contact-afdrukken*). Na ontwikkeling van de laklaag is het maskerpatroon op de plak overgebracht. Aangezien voor de vervaardiging van een transistor of IC meerdere patronen na elkaar moeten worden aangebracht, moeten de verschillende maskers ten opzichte van elkaar worden uitgericht.

Aan dit contact-afdrukken kleeft een aantal bezwaren zoals *maskerbeschadiging* (ten gevolge van het op elkaar drukken van masker en plak kunnen maskers in de produktie slechts een keer of vijf worden gebruikt) en *buigingsverschijnselen* (ten gevolge van het niet overal in perfect contact met elkaar zijn van masker en plak). Bovendien vereist de vervaardiging van het masker vele processtappen.

Hoewel er geen principieel verschil bestaat tussen de vervaardiging van (bijvoorbeeld) een enkele transistor en (bijvoorbeeld) een microprocessor, worden bovengenoemde bezwaren groter als we naar meer complexe circuits gaan. Deze complexiteit komt namelijk, voor wat de technologie betreft, vooral in de lithografie tot uitdrukking, en wel door het gelijktijdig optreden van alsmaar *kleiner wordende minimale afmetingen* (2 à 3 μm), de daarbij behorende *grotere uitrichtnauwkeurigheidseisen* en *het grotere chip-oppervlak* ($>$ 5 x 5 mm^2). De uitrichtnauwkeurigheid moet groter worden omdat het nuttig effect van kleinere afmetingen goeddeels verloren gaat als deze niet gepaard gaan met een nauwkeurige passing van de verschillende lagen. Als vuistregel geldt dat de uitrichtnauwkeurigheid beter dan een kwart van de kleinste afmeting moet zijn.

De categorie circuits met bovengenoemde complexiteit en met meer dan 10^5 transistoren wordt *Very Large Scale Integration* (VLSI) genoemd. De combinatie van bovengenoemde bezwaren maakt het steeds moeilijker om met contactafdrukken te voldoen aan de eisen die aan de vervaardiging van VLSI-circuits worden gesteld, en om die reden wordt er druk gewerkt aan een aantal alternatieve methoden. Een (onvolledig) overzicht van de verschillende mogelijkheden treft U aan in figuur 1. In de loop van dit verhaal zullen verschillende begrippen uit deze figuur nader worden toegelicht. De aard van de oplossingen waaraan wordt gewerkt is soms praktisch (geen contact, minder vervaardigingsstappen), soms fundamenteel (kleinere golflengten) en soms beide.

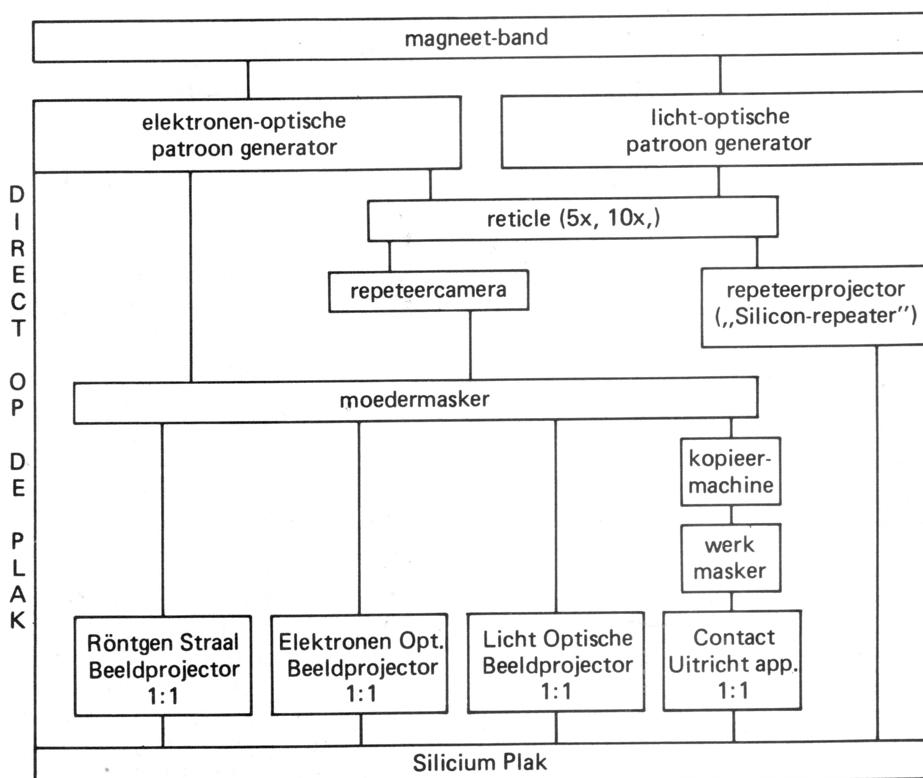


Fig. 1. Schematische voorstelling van verschillende lithografische methoden. De magneetband boven in de figuur bevat in digitale vorm de geometrische gegevens van het ontwerp. Aangegeven is op welke manieren vanaf de magneetband het ontwerp op de Si-plak wordt afgebeeld.

Het is in dit bestek niet doenlijk om al deze verschillende methoden op hun merites te beoordelen; wel is het nuttig om een aantal criteria te noemen die aan deze beoordeling ten grondslag liggen: 1) *kleinst af te beelden afmeting*; 2) *uitrichtnauwkeurigheid*; 3) *repeteermogelijkheid* (in verband met plakvervorming); 4) *automatiseringsaspecten*; 5) *golf- en deeltjesverschijnselen* (staande golven, interferentie, verstrooiing); 6) *focusdiepte*; 7) *uniformiteit van belichting*; 8) *mogelijkheid om voor plakvervorming te corrigeren*; 9) *invloed van plakgrootte*; 10) *onderlinge uitwisselbaarheid* van de machines; 11) *maskervervaardigingsproblemen* en 12) aantal te belichten plakken per uur ofwel: *kosten*! De meeste van deze punten zullen in de loop van dit verhaal nog ter sprake komen.

Als voorbeeld van een methode die duidelijke vooruitgang heeft gebracht, zouden we *licht-optische 1:1-beeldprojectie* (lens of spiegel) willen noemen. Deze heeft als voornaamste voordeel dat masker en plak niet meer met elkaar in contact komen, waardoor geen maskerbeschadiging meer optreedt. Met deze methode kunnen 4-inch-plakken worden belicht tot in het gebied van 2 à 3 μm minimale afmetingen. Mede vanwege de voortreffelijke werking van dit type apparaat is deze methode binnen enkele jaren de industrie-standaard voor de produktie van geavanceerde circuits geworden.

In plaats van met optische technieken wordt tegenwoordig meer en meer het masker vervaardigd door met een elektronenbundel het gewenste patroon in elektrongevoelige lak te schrijven. Deze bundel kan een zeer kleine diameter (ongeveer 0,1 μm) hebben en patronen kunnen zeer nauwkeurig worden aangebracht. De combinatie van elektronenstraalmaskers en optische 1:1-beeldprojectie is dan ook een zeer aantrekkelijke methode. Toch heeft ook deze methode zijn beperkingen (minimale afmetingen, focusdiepte) en we willen daarom twee andere methoden nader toelichten, waarvan de eerste zeker, en de andere misschien, een belangrijk deel van de lithografie ten behoeve van VLSI voor zijn rekening zal nemen.

De Repeteerprojector

Thans wordt het gerepeteerde masker, indien door optische technieken verkregen, vervaardigd door een 5 of 10 maal vergroot circuitpatroon (het reticle) te verkleinen en te repeteren op het uiteindelijke (moeder-)masker. Het ligt nogal voor de hand om dit uiteindelijke masker over te slaan en het reticle direct op de Si-plak te projecteren en te repeteren. Dit nu is het principe van de repeteerprojector (zie figuur 1). Dat deze methode pas sedert korte tijd wordt toegepast heeft te maken met de verschillen die optreden wanneer we in plaats van een masker een Si-plak belichten. Het voornaamste verschil betreft de snelheid, waarmee een gehele plak moet kunnen worden belicht om de methode ook economisch te maken. Ook het feit dat de plak een minder ideaal beeldvlak vormt dan het masker zorgt voor complicaties. Bovendien kan de vorm van een Si-plak in de loop van het vervaardigingsproces van een geïntegreerde schakeling veranderen; hierop komen we later nog terug.

De volgende ontwikkelingen hebben het gebruik van repeteerprojectoren in massaproduktie binnen bereik gebracht.

1) *Lenzen met hoog scheidend vermogen.*
De aan de lenzen te stellen eisen zijn zeer zwaar voor wat betreft de combinatie van scheidend vermogen, veldgrootte en beeldvervorming. Een lens van topkwaliteit van dit moment heeft een Numerieke Apertuur van 0,3 (minimum dimensie 1 à 1,5 μm) bij een beeldveld van 10 x 10 mm^2 en een beeldvervorming van ongeveer 0,1 μm.

2) *Sterke en homogene lichtbronnen.*
Er zijn nu lichtbronnen die in repeteerprojectoren een belichtingstijd van ¼ tot ½ sec geven bij een homogeniteit van ± 2%.

3) *Snelle, met behulp van laser-interferometersystemen gecontroleerde, x-y-positionering.*
De precisie van deze positionering ligt in de orde van ± 0,1 μm en de verplaatsingstijden liggen in de orde van ¼ tot ½ seconde.

4) *Autofocus.*
Omdat het oppervlak van een Si-plak over vele microns in de hoogte kan variëren is het noodzakelijk dat de projectiekolom voor iedere belichting wordt gefocusseerd. Er zijn hiervoor verschillende concepten ontwikkeld.

5) *Uitrichtsystemen.*
Het is in het licht van bovengenoemde minimale afmetingen en eerdere opmerkingen over vereiste uitrichtnauwkeurigheid duidelijk dat een adequaat uitrichtsysteem van het hoogste belang is en dat een automatisch uitrichtsysteem welhaast noodzakelijk is. Het enige automatische uitrichtsysteem dat tot nu toe met succes is toegepast is dat van de repeteerprojector die in het Philips Natuurkundig Laboratorium is ontwikkeld.

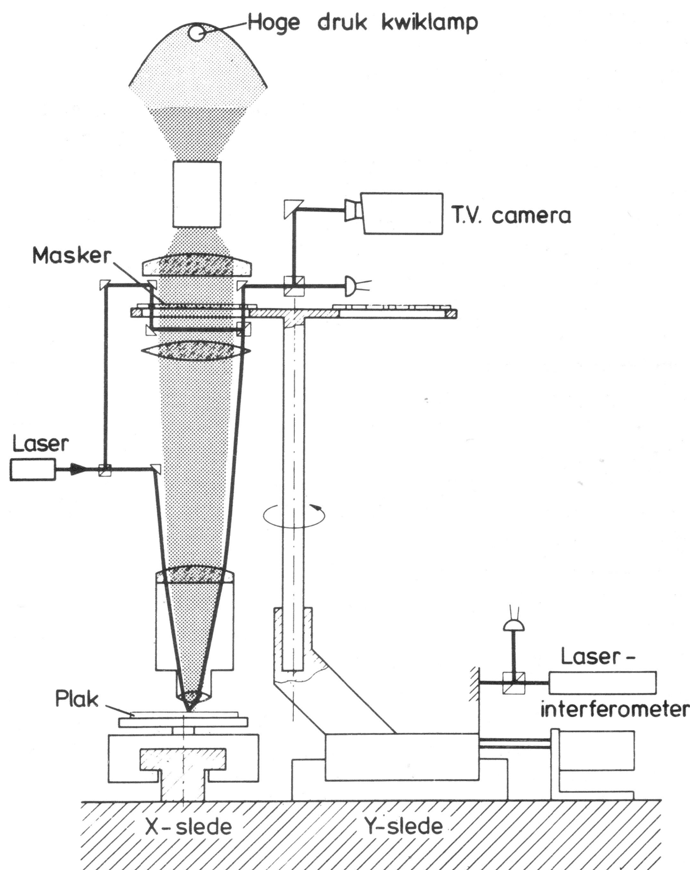
Samenvattend zouden wij de volgende gunstige eigenschappen van de repeteerprojector willen benadrukken: 1. Als masker wordt een reticle gebruikt. Voor de vervaardiging hiervan zijn slechts één of twee stappen nodig en dit zal zeker opbrengstverhogend werken; bovendien is dit reticle eenvoudig te inspecteren en te repareren. 2. Het scheidend vermogen ligt in de orde van 1 tot 1,5 μm. 3. Het systeem is in staat zich aan de vorm van de plak aan te passen. Dit wil zeggen dat de grootte van de plak in eerste benadering niet relevant is en dat voor lokale hoogtevariaties kan worden gecorrigeerd. Bovendien kan bij toepassing van een adequaat uitrichtsysteem voor laterale plakvervorming (waarover later) worden gecorrigeerd, hetzij door de grootte van de repeteerstap te variëren, hetzij door op meerdere plaatsen uit te richten.

Het is vanwege de combinatie van deze voordelen dat wij ervan overtuigd zijn dat er voor de repeteerprojector een zeer ruime plaats zal zijn ingeruimd in het VLSI-tijdperk.

De Silicon-Repeater

Als voorbeeld van een repeteerprojector, zouden wij enige aandacht willen besteden aan de Philips repeteerprojector, de Silicon-Repeater (Si-Re).
Figuur 2 toont een schematische voorstelling van de Si-Re, dit apparaat werkt met een verkleiningsfactor van 5.

Er is uitgegaan van twee hydrostatisch gelagerde sleden die door hydraulische lineaire motoren langs onderling loodrechte banen bewogen kunnen worden. De ene slede draagt, op een speciale instelbare houder, de met fotolak bedekte Si-plak, de andere draagt de projectiekolom. Twee laser-interferometersystemen maken het mogelijk om de positie van beide sleden met een nauwkeurigheid beter dan ± 0,1 μm te meten. De projectiekolom bevat als lichtbron een superhogedrukkwiklamp. Een elliptische spiegel achter de lamp en een speciaal condensorsysteem zorgen voor een voldoend sterke en homogene verlichting van het te projecteren masker. Het lenzensysteem is telecentrisch, dat wil zeggen dat bij projecteren op een oneffen beeldveld de vergroting niet varieert. Een bijzonder sterk punt van deze Si-Re is het automatische uitrichtsysteem, dat inmiddels is beschreven[1]. Ten behoeve van dit uitrichtsysteem bevatten zowel masker als plak twee uitrichtkenmerken bestaande uit rasters in de x- en y-richting. De rasters op de plak worden door een laserbundel (links in de figuur) belicht en door de projectielens afgebeeld op de overeenkomstige rasters op het masker. Voor verdere details zij naar genoemde publikatie verwezen. Met behulp van de informatie die tijdens de automatische uitrichtcyclus wordt verkregen is het in principe mogelijk om de door het laser-interferometer-systeem gecontroleerde repeteerstap-grootte automatisch aan te passen en zo voor eventuele plakvervorming te corrigeren. Als bijvoorbeeld de plak groter is geworden gedurende het IC-vervaardigingsproces, dan moet een grotere slede-verplaatsing worden uitgevoerd. Het is ook mogelijk



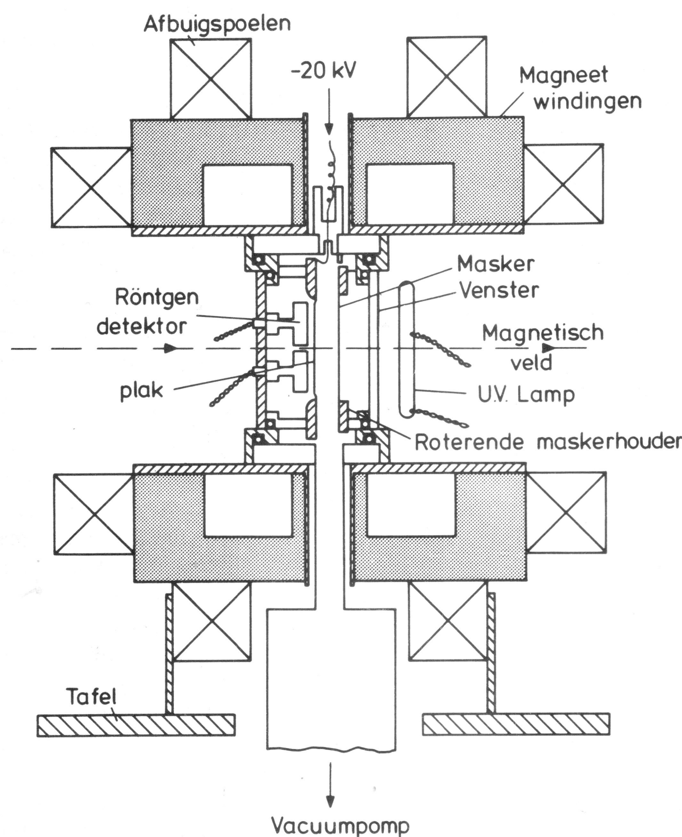
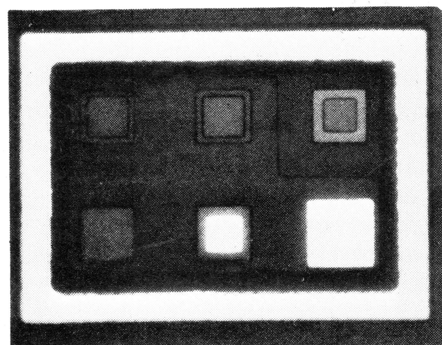


Fig. 2. Schema van de Silicon-Repeater.

om meerdere uitrichtkenmerken aan te brengen, en dus navenant vaker uit te richten. De uitrichtnauwkeurigheid van dit systeem is in de orde van ± 0,1 μm. In de praktijk vinden we een 3σ-waarde van ± 0,25 μm. We hebben nu enige jaren ervaring opgedaan met dit apparaat en op grond van deze ervaring wordt nu gewerkt aan een tweede generatie, met als voornaamste verbetering een hogere produktiesnelheid. Nog enkele getallen, die een indruk geven van de prestaties van de huidige Si-Re: maximaal maskerbeeld: 50 x 50 mm^2 ; beeldveld bij maximale scherpte: 7 x 7 mm^2 ; kleinste detail in schakeling: 1,5 tot 2 μm; belichtingstijd per beeldveld 0,5 s; verplaatsingstijd: 0,5 seconde.

Fig. 3. Foto, waarop de automatische uitrichtnauwkeurigheid van de Silicon-Repeater te zien valt. Het verschil in grootte tussen twee op elkaar passende vierkantjes, die in twee verschillende belichtingen zijn aangebracht, is steeds 1 à 2 μm. Het betreft hier een 5-masker-proces.



Figuur 3 laat een voorbeeld zien van de uitrichtnauwkeurigheid die wordt behaald bij de vervaardiging van een IC. De paarsgewijze in elkaar passende vierkantjes zijn hulpfiguren die met de verschillende maskerlagen op elkaar worden gelegd en bieden de mogelijkheid om de uitrichtnauwkeurigheid visueel te controleren; het verschil in grootte tussen twee op elkaar passende vierkantjes is 1 à 2 μm.

De elektronenoptische beeldprojector[2]

Principe

De methode berust erop dat het patroon van een in vacuüm geplaatst masker door middel van elektronen via 1:1-projectie op (bijvoorbeeld) een siliciumplak wordt overgebracht. Hiertoe wordt een chroommasker, in essentie identiek aan een masker zoals dat voor conventionele foto-lithografie wordt gebruikt, bedekt met een *fotokathode* (ongeveer 200 Å cesiumjodide). Dit masker wordt van achteren beschenen met ultraviolet licht, zodat elektronen worden geëmitteerd in al die gebieden die niet door Cr zijn gemaskeerd. Deze elektronen worden door een elektrisch veld versneld naar een met een elektrongevoelige laklaag (PMMA, polymethylmetacrylaat) bedekte Si-plak en gefocusseerd door een magnetisch veld.

Fig. 4. Schema van de elektronenoptische beeldprojector.

Omdat we met UV-licht van 184,9 nm werken, moet het maskersubstraatmateriaal kwarts zijn. Het uitrichten van de opeenvolgende lagen gebeurt *automatisch* met een uitrichtnauwkeurigheid van ± 0,1 μm. Translatiefouten worden gecorrigeerd door het aanpassen van twee, als afbuigspoelen fungerende, hulpmagneetvelden, die loodrecht op elkaar en op het focusseerveld staan; rotatiefouten worden mechanisch geëlimineerd door het masker te roteren. In figuur 4 is een schematische doorsnede van de projector te zien.

Hierna wordt op enkele deelaspecten van dit systeem nader ingegaan.

Fotokathode

CsI is in een elektronenoptische beeldprojector een geschikte fotokathode. Deze stof voldoet aan alle hoofdeisen voor gebruik in een machine voor massaproduktie: eenvoudige bereiding en verwijdering, stabiel gedrag in droge lucht, betrouwbare werking in een betrekkelijk slecht vacuüm (10 mPa) en een stabiele fotostroom met een stroomdichtheid van tenminste 1 μA/cm^2 .

De combinatie van deze gunstige eigenschappen zorgt ervoor dat de laag CsI in de machine zelf aangebracht kan worden, ongeveer 50 belichtingen kan verrichten en bij gebruik van PMMA (gevoeligheid

ongeveer 100 $\mu C/cm^2$) een belichtingstijd in de orde van 10 seconden oplevert. Het masker zelf lijdt niet onder het gebruik en heeft, in principe, het eeuwige leven.

De gebruikte combinatie van UV-fotonen (lagedrukkwiklamp met een golflengte van 184,9 nm, foto-energie = 6,7 eV) en CsI als fotokathode (uittree-arbeid = 6,5 eV) levert een belangrijk voordeel op. Dit voordeel is dat de energie van de geëmitteerde elektronen maximaal ongeveer 0,5 eV bedraagt. Hierdoor is er slechts een geringe spreiding in deze beginenergie en hebben we in het beeld geen last van het elektronenoptisch equivalent van chromatische aberratie en dus van verlies aan scherpte en scherpte-diepte.

Beeldvorming, scheidend vermogen en beeldvervorming

De elektronen, die door de fotokathode worden geëmitteerd, worden naar de Si-plak (de anode!) toe versneld door een elektrische spanning van ongeveer 20 kV en gefocusseerd door een magnetisch veld van ongeveer 80 kA/m. De plaats van het focus hangt af van de sterkte van het magnetisch veld; teneinde eventuele vertekeningen in het beeld kleiner dan 0,1 μm te houden, mogen afwijkingen in de homogeniteit van het veld ten hoogste van de orde van grootte van 10^{-5} zijn, over het gehele oppervlak van de plak. Met behulp van een samengestelde solenoïde is een magnetisch veld verkregen, waarvan de homogeniteit over een gebied van 50 mm diameter beter dan 3 op 10^5 is.

De elektronen worden niet alleen door het magnetisch veld gefocusseerd, maar in feite door de combinatie van het magnetisch veld en het elektrisch veld. Het focusseren van het systeem geschiedt in de praktijk dan ook door het veranderen van de versnelspanning. De combinatie van elektrisch en magnetisch veld moet ervoor zorgen dat de elektronen, als ze de afstand masker-plak (in dit geval 16 mm) hebben afgelegd, precies een hele omwenteling rond de magnetische veldlijnen hebben gemaakt.

Als de machine in focus is, is aan de volgende formule voldaan:

$$\cos \alpha \sqrt{v_0/eV} - 1 + (\pi/Bd) \cdot \sqrt{2\,mV/e} = 0$$

waarin:
- α = uittree-hoek van de geëmitteerde elektronen
- v_0 = uittree-energie (ongeveer 0, 2 eV)
- e = elementaire lading
- V = versnelspanning
- B = magnetische inductie
- d = afstand tussen anode en kathode
- m = massa van het elektron.

Het *scheidend vermogen* van dit systeem wordt niet zozeer bepaald door elektronenoptische factoren (homogeniteit magneetveld, uittree-energie van de elektronen – als we ons zouden beperken tot deze zuiver elektronenoptische overwegingen, dan zou dit leiden tot een scheidend vermogen van 0,01 μm!), als wel door factoren als lakdikte, onvlakheid van de plak, elektronenverstrooiing in lak en substraat, dikte van het CsI, randscherpte van het Cr-masker enzovoort. Hoewel het niet mogelijk is om een exact getal te geven voor het gezamenlijke effect van deze factoren, is het uit ervaring duidelijk geworden dat het scheidend vermogen van de 'opvallende' elektronen in de orde van enkele tienden van microns ligt. Eén van de gunstige aspecten van dit systeem is de grote scherptediepte: als we een randonscherpte toelaten van 0,1 μm dan is die scherptediepte 100 μm en dit betekent onder andere dat de onvlakheid van de plak ons in eerste benadering niet erg hindert.

De term 'opvallende' elektronen is hierboven met opzet gekozen om aan te geven dat het scheidend vermogen van dit systeem (evenals dit het geval is met alle elektronenoptische systemen) veel meer wordt bepaald door wat er met de elektronen gebeurt als ze de Si-plak met een eindige laklaagdikte hebben bereikt. In de lak, tegen en in het substraat worden de elektronen namelijk verstrooid en gedeeltelijk teruggekaatst. En dit kan, afhankelijk van lakdikte en substraatmateriaal, aanleiding geven tot een significante lijnverbreding. Bovendien, anders dan in elektronenoptische patroongeneratoren, worden elektronen die na terugkaatsing uit de laklaag te voorschijn komen, door het daar heersende elektromagnetische veld weer naar de plak teruggebogen en geven daar aanleiding tot (ongewenste) belichting. Dit proces kan zich over verschillende mm's uitstrekken en veroorzaakt verdere randvertekening. Hier dient nog verder onderzoek te worden verricht. Niettemin lijkt het, mede gezien onze ervaring, verantwoord om te stellen dat dit systeem in staat is om in het submicrongebied te werken.

Hierboven is al aangegeven dat *elektronenverstrooiing* in lak en substraat aanleiding kan geven tot beeldvervorming. Er zijn ook andere faktoren die beeldvervorming veroorzaken. *Beeldvervorming* is een gecompliceerd onderwerp en moet worden onderverdeeld in reproduceerbare en niet-reproduceerbare beeldvervorming. Het zal duidelijk zijn dat de niet-reproduceerbare beeldvervorming, dat wil zeggen die beeldvervorming die van belichting tot belichting ofwel van laag tot laag verschillend is, zeer ernstige consequenties kan hebben bij de vervaardiging van geïntegreerde schakelingen, waarbij 6 tot 10 lagen met een grote nauwkeurigheid op elkaar moeten passen.

Zoals we hierboven al gezien hebben, kan het magnetisch veld voldoende uniform worden gemaakt om de vervorming beneden de 0,1 μm te houden, maar met het elektrische veld ligt dat heel anders (de plak is de anode!). Het elektrische veld wordt bepaald door masker, plak en de houders hiervan: er zijn onvermijdelijke randeffecten en discontinuïteiten die het veld verstoren. In de praktijk vinden we meerdere microns vertekening aan de rand van de plak, maar deze vertekening is reproduceerbaar en gaat niet ten koste van de geëiste passingsnauwkeurigheid. Een ernstige oorzaak van beeldvervorming is verandering van plakkromming. Gedurende het vervaardigen van geïntegreerde schakelingen verandert de onvlakheid van de Si-plak ten gevolge van oxidaties, diffusies en gedeponeerde lagen. Uit berekeningen en experimenten is gebleken dat op 2-inch-plakken ongeveer een twintigste van de verandering van de plakkromming als een dwarsverplaatsing van het beeld zichtbaar wordt. Aangezien deze verandering enkele tientallen microns kan bedragen is het duidelijk dat dit tot ontoelaatbare passingsfouten leidt. De voor de hand liggende oplossing is om de plak te dwingen vlak te liggen. Dit kan niet gebeuren, zoals bij optische lithografie, door een vacuümaanzuigplaat, maar aangetoond is dat we gebruik kunnen maken van de hoogspanning die al in het systeem aanwezig is om de plak vlak te houden door elektro-statische aantrekkingskracht.

Door de plak vlak te houden vermijden we wel deze vorm van beeldvervorming, maar heffen we nog niet de spanning in de plak op. Oxidaties, gedeponeerde lagen en diffusies veroorzaken namelijk tevens spanningen die, in eerste benadering, lineaire uitzettingen of inkrimpingen van de plak veroorzaken die met het krapper worden van de uitrichttoleranties te groot zijn om te verwaarlozen. Ook temperatuurverschillen tussen de opeenvolgende belichtings- en uitrichtstappen veroorzaken dit effect, dat overigens alle 1:1-belichtingssystemen hindert. De elektronenoptische beeldprojector kan dit effect echter corrigeren. Er is aangetoond dat een kleine aanpassing van het magneetveld een geringe vergroting of verkleining van de 1:1-maatstaf kan bewerkstelligen. Om dit automatisch te regelen, gebruiken we de informatie die reeds beschikbaar is in de röntgensignalen die vanaf de uitrichtkenmerken komen. Hiermee zijn we op het aspect van het automatisch uitrichten gekomen.

Het automatisch uitrichten

Het werkgebied van de elektronenoptische beeldprojector is de produktie van patronen met minimale afmetingen in de orde van 1 μm. Dit betekent dat de gewenste uitrichtnauwkeurigheid in de orde van 0,25 μm ligt. In een machine die is bestemd voor massaproduktie moet men

een uitrichtsysteem hebben dat behalve nauwkeurig, zowel snel als automatisch werkt. Dientengevolge is er een methode ontwikkeld, die gebruikmaakt van de remstraling die ontstaat wanneer elektronen een zwaar element treffen. In deze methode wordt een raster van markeringen (van tantaaloxyde) op de Si-plak aangebracht en een identiek patroon in het masker. Wanneer het beeld van het raster op het masker precies samenvalt met het raster op de plak, vallen al de elektronen van het masker-raster op het tantaaloxyde; wanneer er echter een fout in het uitrichten zit, zullen sommige elektronen op het silicium terechtkomen, dat een kleiner rendement heeft voor de opwekking van röntgenstralen. Zo zal dus bij perfect uitrichten de opbrengst aan röntgenstralen maximaal zijn. Achter de plak gemonteerde detectors zetten de röntgenstraling die door de plak heengekomen is om in elektrische signalen. De instelling op maximale intensiteit – d.i. het uitrichten – geschiedt automatisch. De door de detectors geproduceerde signalen worden fasegevoelig gedetecteerd, wat tot eenvoudige elektronische schakelingen leidt. Er wordt zowel aan de linker- als aan de rechterzijde van de plak/het masker een raster van markeringen gebruikt, hierdoor kunnen niet alleen translatie-, maar ook rotatiefouten worden gedetecteerd en gecorrigeerd. De translatiecorrecties worden uitgevoerd door middel van de afbuigspoelen, maar de rotatiecorrecties kunnen niet op dezelfde wijze worden uitgevoerd; dat zou een onaanvaardbare vertekening van het beeld opleveren. In plaats daarvan wordt de maskerhouder gedraaid. Het uitrichten kan, door gebruik van snelle halfgeleiderdetectors, binnen de eerste 20% van de totale bestralingstijd plaatsvinden. De rest van de plak wordt gedurende het uitrichten eveneens bestraald, maar dit komt, omdat de gebruikte elektronenlak PMMA een vrij hoog contrast heeft, niet tot uitdrukking in het ontwikkelde beeld. Daarmee hebben we het grote voordeel dat een aparte belichting voor de uitrichtkenmerken niet nodig is.

Het doel van deze uitrichtmethode is het behalen van een uitrichtnauwkeurigheid van ± 0,1 μm ter plekke van de markeringen. Deze nauwkeurigheid wordt bereikt als we uitgaan van ongeproceste plakken en de temperatuur goed in de hand hebben. In de praktijk echter, tijdens de vervaardiging van geïntegreerde schakelingen, is deze nauwkeurigheid, mede ten gevolge van de in de vorige paragraaf vermelde plakvervorming, wel iets minder, maar is toch vrijwel *altijd beter dan ± 0,25 μm*. Dit betreft uiteraard de uitrichtnauwkeurigheid ter plekke van de markeringen: elders op de plak kan de passingsnauwkeurigheid door een aantal eerder vermelde oorzaken (verandering van plakkromming, plakvervorming) minder zijn.

Resultaten en conclusies

Er is nu enige jaren ervaring met een aantal van deze elektronenoptische beeldprojectoren, er zijn zowel bipolaire als MOST-circuits mee gemaakt en samenvattend kan worden gesteld dat de machines aan hun doelstelling, het automatisch belichten en uitrichten van structuren tot in het submicrongebied, beantwoorden. Als sterke punten gelden dus het hoge scheidend vermogen en de hoge automatische uitrichtnauwkeurigheid. Bovendien is het systeem relatief goedkoop. Als nadelen moeten worden genoemd de gevoeligheid voor plakvervorming, de effecten ten gevolge van elektronenverstrooiing en de noodzaak van het invoeren van tantaaluitrichtkenmerken.

De huidige machines hebben een cyclustijd van ongeveer 3 minuten, maar het moet niet moeilijk zijn om dit in een produktiemachine aanmerkelijk te verkorten, aangezien van deze 3 minuten slechts ongeveer 40 seconden wordt belicht. Het is nog te vroeg om uitspraken over procesopbrengst te doen.

Wel is duidelijk dat de spectaculaire vooruitgang in de laatste jaren van optische belichtings- en uitrichtapparatuur de behoefte aan deze elektronenoptische beeldprojector naar de toekomst verschoven heeft. Er lijken echter geen fundamentele redenen te zijn die dit systeem beletten om in het submicrongebied een levensvatbare en economisch haalbare methode voor massaproduktie te zijn.

Enige slotopmerkingen

Niet alle methoden die in het schema van figuur 1 voorkomen zijn even uitgebreid besproken.
Er is bijvoorbeeld nog niets gezegd over röntgenstraallithografie, een methode die voor het submicrongebied interessant kan worden omdat er noch buiging, noch verstrooiing optreedt.
Daarnaast zijn er nog methoden in ontwikkeling, die op het schema niet voorkomen, zoals de elektronenoptische repeteerprojector en, meer recent, ionen- en laserstraallithografie.

Er is in dit artikel niet gepoogd om de verschillende lithografische methoden te behandelen met een uitgebreidheid die evenredig is aan hun belang, voor zover dat althans te overzien zou zijn. We hebben een overzicht gegeven van methoden die ons in het VLSI-tijdperk ter beschikking staan. Daarbij hebben we speciale aandacht besteed aan licht-optische repeteerprojectie, als voorbeeld van een methode die zeer zeker een belangrijk deel van de lithografie voor zijn rekening zal nemen en aan elektronenoptische beeldprojectie, als voorbeeld van een methode die, hoewel zeer interessant, veel speculatiever is voor wat betreft zijn feitelijke toepassing.

Anderzijds is, gezien tegen het belang van de techniek, te weinig aandacht besteed aan de methode die elektronenstraalmaskers en licht-optische 1:1-beeldprojectie combineert. Met deze methode zal in de komende jaren de grote meerderheid van de geavanceerde circuits worden vervaardigd. Dit te meer omdat er ontwikkelingen gaande zijn om deze projectie in het verre ultraviolet te laten plaatsvinden, waardoor het mogelijk zal zijn om nog kleinere afmetingen af te beelden.

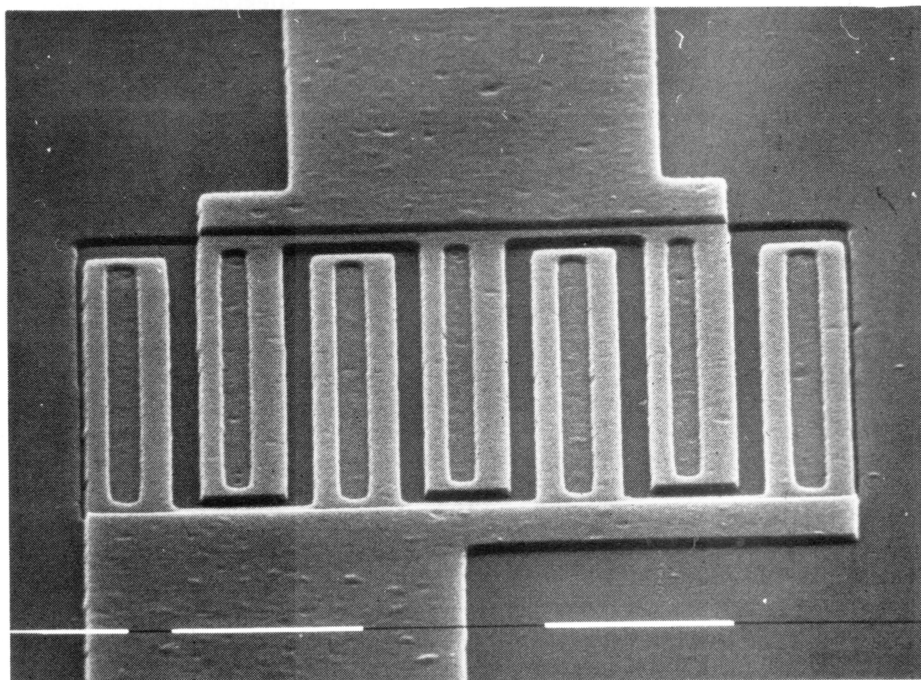


Fig. 5. SEM-opname van een bipolaire hoogfrequent transistor. De verschillende (4) lagen van deze transistor zijn gedefinieerd door met behulp van een elektronenoptische patroongenerator direct op de plak te schrijven. Elk van de witte streepjes stelt 10 micron voor.

De algemeen aanvaarde visie op het VLSI-tijdperk voor wat betreft de lithografie is dan ook dat tot in het 1-microngebied licht-optische 1:1-beeldprojectie en licht-optische repeteerprojectie de feitelijk toegepaste methoden zullen zijn.

Wat er daarna, in het submicrongebied, zal gebeuren is veel minder duidelijk. Zowel röntgenstraallithografie als elektronenoptische beeldprojectie maken daar een kans, het vereist echter een apart artikel om de verschillende fysische problemen, moeilijkheden en mogelijkheden van deze (en andere!) methoden evenwichtig tegen elkaar af te wegen. Er is voor dit submicrongebied nog minstens één andere kandidaat en dat is de methode die op het schema van figuur 1 als *'direct op de plak'* staat aangegeven.

We willen dit artikel besluiten met een resultaat van deze methode. In dit geval wordt het patroon, dat in digitale vorm op de magneetband staat, direct door de elektronenoptische patroongenerator op de plak geschreven. Het is duidelijk dat deze methode vele voordelen biedt; het grootste nadeel is dat bij de huidige schrijftijden deze manier van patroondefinitie veel te duur is om voor gebruik in massaproduktie in aanmerking te komen.

Figuur 5 toont een SEM-opname van een bipolaire hoogfrequent-transistor met submicronbasis en emittercontacten; naast de scherpte van het patroon valt vooral de perfecte (automatische) uitrichtnauwkeurigheid op. Ook voor deze methode zal er in de toekomst zeker een plaats zijn.

Referenties

1. G. Bouwhuis, S. Wittekoek: *Automatic Alignment System for Optical Projection Printing.* IEEE Tr. Electron Devices **26** (1979) 723.
2. J.P. Scott: *An Electron Image Projector with Automatic Alignment.* IEEE Tr. Electron Devices **22** (1975) 409.

Schottky-diode versus pn-diode

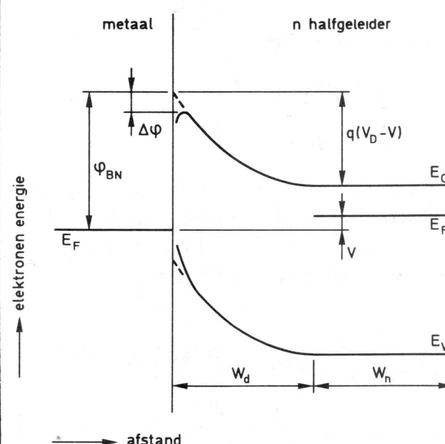
M. Kleefstra

Prof. dr. M. Kleefstra (49), sinds 1 januari 1980 hoogleraar, studeerde scheikunde te Groningen en promoveerde aldaar in 1961 op een proefschrift over kleurcentra. Van 1960 tot 1969 werkte hij aan elektronische componenten bij Thomson-CSF in Frankrijk. Sindsdien is hij verbonden aan de Afdeling der Elektrotechniek van de Technische Hogeschool te Delft, vanaf 1977 als lector in de elektrotechnische materialen. Hij houdt zich daar bezig met Schottky-dioden en met CCD's die door pn-dioden worden aangestuurd.

De stroom-spanningskarakteristieken van Schottky-dioden en die van pn-dioden in matig en homogeen gedoteerde halfgeleiders worden gegeven door vrijwel dezelfde formules. Bij de gebruikelijke pn-dioden wordt de effectieve snelheid van de ladingsdragers begrensd door diffusie van ladingsdragers. In Schottky-dioden wordt de snelheid van de meerderheidsladingsdragers bepaald door de thermionische emissie van die ladingsdragers en moet de verlaging van de potentiaalberg door de beeldkracht worden meegenomen. Of het ene type diode te verkiezen is boven het andere type hangt af van de toepassing en van de materiaalkeuze.

In geïntegreerde schakelingen komen alom metaal-halfgeleider- en p-n-overgangen voor. Bij matig gedoteerde halfgeleiders werken deze overgangen als gelijkrichter en spreekt men van respectievelijk Schottky- en pn-diode. Beide typen dioden worden vaak apart en onvolledig behandeld[1,2,3], zodat de overeenkomsten en verschillen moeilijk te zien zijn. De Schottky-diode alleen wordt duidelijk door Rhoderick[4] behandeld, de pn-diode door Grove[5].
Hier zal worden aangetoond dat voor dioden met matige dotering, van 10^{21} – 10^{24} m^{-3} in silicium, de klassieke formules voor het *I-V*-verband dusdanig kunnen worden geschreven dat overeenkomsten en verschillen zo te zien zijn.
Voor dioden met hogere doteringen zijn de formules zeer verschillend en deze zullen hier buiten beschouwing blijven.
Dioden met matige dotering worden gebruikt in gelijkrichters, zonnecellen, basis-collectorovergangen in transistoren en in CCD's.
De beschouwingen zullen worden beperkt tot het gedrag rond kamertemperatuur, de dotering van de halfgeleiders wordt homogeen genomen, de rand van de ontruimingslaag wordt oneindig scherp genomen en de weerstand van contacten op de niet-diode-kant van de halfgeleider wordt steeds verwaarloosd. De gebruikte symbolen worden in tabel I toegelicht en voor de halfgeleiders silicium en galliumarsenide worden waarden voor de parameters aangegeven.
Het *I-V*-verband zal met behulp van drie mechanismen worden berekend:
– de recombinatie en generatie van ladingsdragers in de ontruimingslaag,
– de thermionische emissie van ladingsdragers over een potentiaalberg,
– de diffusie van ladingsdragers als minderheden in een stuk halfgeleider nadat de ladingsdragers de potentiaalberg hebben 'genomen' dankzij de thermionische emissie. Met behulp van de Fermi-Dirac-statistiek en de diffusievergelijking $J_n = qD\,dn/dx$ zijn de formules voor het *I-V*-verband af te leiden. Pas na numeriek uitrekenen kan worden beslist welk mechanisme overheerst bij de gekozen parameters.

Fig. 1. Energieschema van de Schottky-diode.



De I-V-karakteristiek van de Schottky-diode

De toegelaten energieniveaus van een Schottky-diode op een n-halfgeleider zijn in figuur 1 aangegeven. De bandbuiging tengevolge van de spiegelbeeldkracht is voor E_c en E_v aangegeven. In figuur 1 wordt het potentiaalverschil $V_D - V$ opgebouwd door de positieve ruimtelading in de ontruimingslaag met dikte W_d. Door de werking van de spiegelbeeldkracht wordt de potentiaalberg verlaagd met

symbool	definitie	eenheid	waarde(n) Si		GaAs
D_n	diffusiecoëfficiënt van elektronen	m^2/s	3×10^{-3}		2×10^{-2}
D_p	diffusiecoëfficiënt van gaten	″	1×10^{-3}		8×10^{-4}
E_c	onderkant van de geleidingsband	eV			
E_F	Fermi-niveau	″			
E_v	bovenkant van de valentieband	″			
E_g	breedte van de verboden band	″	1,12		1,43
J	stroomdichtheid	$A\,m^{-2}$			
L_n	difussielengte van de elektronen	m	$L_n = \sqrt{D_n\,\tau}$		
L_p	difussielengte van de gaten	″	$L_p = \sqrt{D_p\,\tau}$		
T	absolute temperatuur	K	300		300
N_A	concentratie van acceptoratomen	m^{-3}	$10^{21} - 10^{24}$	(a)	$10^{21} - 10^{24}$
N_D	concentratie van donoratomen	″	″	″	″
V	aangelegde voorwaartse spanning	V			
V_D	diffusiespanning	″	$0 - 1$	″	$0 - 1$
W_d	dikte van de depletielaag	m	$0 - 10^{-6}$	″	$0 - 10^{-6}$
W_n	dikte van de geleidende n-laag	″	$0 - 10^{-5}$	″	$0 - 10^{-5}$
W_p	dikte van de geleidende p-laag	″	$0 - 10^{-5}$	″	$0 - 10^{-5}$
k	constante van Boltzmann	Joule/K	$1{,}38 \times 10^{-23}$		$1{,}38 \times 10^{-23}$
n	concentratie van elektronen	m^{-3}	$0 - 10^{24}$	″	$0 - 10^{24}$
n_i^2	$= n \times p$ bij evenwicht	m^{-6}	2×10^{32}		1×10^{26}
p	concentratie van gaten	m^{-3}	$0 - 10^{24}$	(a)	$0 - 10^{24}$
v_{th}	thermische snelheid ladingsdragers	$m \cdot s^{-1}$	1×10^5		1×10^5
ϵ	permittiviteit	$F \cdot m$	$1{,}0 \times 10^{-10}$		$1{,}0 \times 10^{-10}$
τ	levensduur van ladingsdragers	s	$10^{-10} - 10^{-4}$	(a)	$\approx 10^{-8}$ (b)

TABEL I. Gebruikte symbolen. (a) min of meer instelbaar. (b) zie referentie 6.

$$\Delta\varphi = \frac{1}{\pi^{1/2}} \left(\frac{q}{2\epsilon}\right)^{3/4} [(V_D - V)N_D]^{1/4}.$$

Wordt de stroom in deze diode bepaald door generatie of recombinatie van ladingsdragers in de ontruimingslaag, dan is een benaderde formule voor de stroomdichtheid

$$J_{rg} = q \cdot W_d \cdot \frac{n_i}{2\tau} (e^{qV/2kT} - 1). \qquad (1)$$

Hierbij wordt de dikte van de ontruimingslaag gegeven door

$$W_d = \left(\frac{2\epsilon(V_D - V)}{qN_D}\right)^{1/2}.$$

Kan de generatie-recombinatie in de ontruimingslaag verwaarloosd worden, dan wordt de stroom van elektronen bepaald door hun thermionische emissie over de potentiaalberg en wordt verkregen

$$J_n = q \cdot \frac{v_{th}}{\sqrt{6\pi}} \cdot N_D \cdot e^{-q(V_D - \Delta\varphi)/kT} (e^{qV/kT} - 1). \qquad (2)$$

De gatenstroom wordt bepaald door diffusie van gaten in het n-gebied en hierin treedt de grootheid W_n op indien $L_p >> W_n$, maar L_p treedt op wanneer $L_p << W_n$. Voor de diode met eindige lengte, dat wil zeggen $L_p >> W_n$ is de gatenstroomdichtheid

$$J_p = q \cdot \frac{D_p \cdot n_i^2}{W_n \cdot D_N} (e^{qV/kT} - 1). \qquad (3a)$$

Voor het geval $L_p << W_n$ geldt:

$$J_p = q \cdot \frac{D_p \cdot n_i^2}{L_n \cdot D_N} (e^{qV/kT} - 1). \qquad (3b)$$

De totale stroomdichtheid is òf J_{rg} òf $J_n + J_p$. Van die twee moet de grootste waarde bij een bepaalde spanning worden genomen. Alleen bij het punt waar $J_{rg} \approx J_n + J_p$ zijn deze formules niet te gebruiken.

De I-V-karakteristiek van de pn-diode

Het bandenschema van de pn-diode is in figuur 2 aangegeven. Omdat hier geen overgang perfecte geleider-ontruimde halfgeleider bestaat, treedt er geen beeldkrachtverlaging op. De stroomdichtheid te danken aan generatie-recombinatie in de ontruimingslaag is hier weer

$$J_{rg} = q \cdot W_d \cdot \frac{n_i}{2\tau} (e^{qV/2kT} - 1). \qquad (4)$$

Hierbij is ook weer

$$W_d = \left(\frac{2\epsilon(V_D - V)}{qN_D}\right)^{1/2}.$$

Bij verwaarlozing van de stroom volgens (4) wordt de elektronenstroom bepaald door de diffusie van elektronen in het p-gebied volgens de formules:

$$J_n = q \cdot \frac{D_n}{W_p} \cdot N_D \cdot e^{-qV_D/kT} (e^{qV/kT} - 1)$$

$$\text{voor } L_n >> W_p \qquad (5a)$$

en

$$J_n = q \cdot \frac{D_n}{L_n} \cdot N_D \cdot e^{-qV_D/kT} (e^{qV/kT} - 1)$$

$$\text{voor } L_n << W_p. \qquad (5b)$$

Het is gebruikelijk deze formules te schrijven als bijvoorbeeld

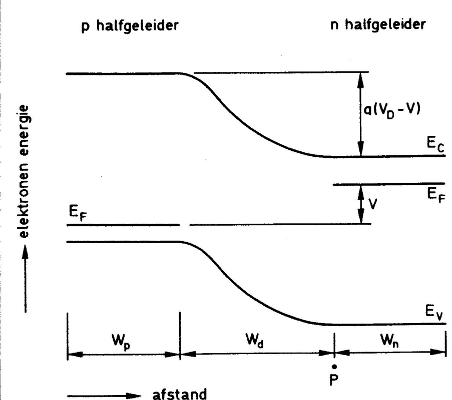
$$J_n = q \cdot \frac{D_n \cdot n_i^2}{W_p \cdot N_A} (e^{qV/kT} - 1),$$

maar hierbij is vergelijken met formule (2) moeilijker. De gatenstroomdichtheid is weer voor de diode met eindige lengte

$$J_p = q \cdot \frac{D_p \cdot n_i^2}{W_n \cdot N_D} (e^{qV/kT} - 1) \qquad (6)$$

Wordt een zeer grote stroom door een

Fig. 2. Energieschema van een pn-diode.



pn-diode geleid waarin voornamelijk de gaten de stroom verzorgen dan moet, vanwege de elektroneutraliteit op het punt P de elektronenconcentratie toenemen en zal aldaar gelden $p = n = n_i$ $e^{qV/2kT}$. De gatenstroomdichtheid wordt dan

$$J_p = q \cdot \frac{D_p \cdot n_i^2}{W_n \cdot n_i} e^{qV/2kT},$$

en verandert dus minder snel met de aangelegde spanning dan in formule (6). De totale stroom door de diode is weer J_{rg} of $J_n + J_p$ waarbij de grootste waarde moet worden genomen.

Overeenkomsten en verschillen

Bij pn-dioden is de diffusiespanning eenvoudig te berekenen uit de doteringen in de p- en de n-gebieden. Omdat er slechts één fase aanwezig is geldt: diffusiespanning = verschil in uittree-arbeiden = contactpotentiaal. Bij een metaal-halfgeleiderovergang is er sprake van een contact tussen twee fasen. De grootte van de diffusiespanning hangt zowel af van het verschil in uittree-arbeiden als van het aantal oppervlaktetoestanden (Tamm-toestanden) op de halfgeleider en de ligging van de oppervlaktetoestanden ten opzichte van de verboden band. De variatie die in de diffusiespanning mogelijk is, is bij beide typen dioden beperkt en globaal kan gesteld worden dat een Schottky-diode wat dit betreft nogal overeenstemt met een p^+n-diode waarbij door variëren van hetzij het metaal, hetzij de dotering in het p^+-gebied de diffusiespanning kan worden veranderd van ongeveer 0,4 tot ongeveer 1,0 volt.

In de *IV*-karakteristieken zijn de meeste stroomcomponenten voor beide typen dioden dezelfde, nl. de stroom J_{rg} en de gatenstroom J_p. Bij de elektronenstroom geeft de Schottky-diode een hogere stroomdichtheid om twee redenen:

– de ingebouwde potentiaalsprong V_D wordt door de beeldkracht met $\Delta\varphi$ verlaagd wat vooral bij sperren invloed heeft,

– bij de Schottky-diode hebben de elektronen in figuur 1 naar links de hoogst mogelijke snelheid, nl. $v_{th}/\sqrt{6\pi}$ terwijl in het p-gebied van de pn-diode de elektronen door diffusie hun gemiddelde effectieve snelheid tot D_n/W_p gereduceerd zien.

Bij gelijke V_D en doteringen heeft de Schottky-diode een veel hogere verzadigingsstroom dan de gebruikelijke pn-diode maar een pn-diode met zeer kleine W_p zal een bijna evengrote verzadigingsstroom bezitten. Vanwege de hogere verzadigingsstroom kan bij Schottky-diodes de stroom J_{rg} vaak worden verwaarloosd en bovendien treedt ook geen term exp $(qV/2kT)$ in plaats van exp (qV/kT) op bij hogere stroomdichtheden.

De opslag van minderheidsladingsdragers in pn-dioden verhindert het zeer snelle schakelen van deze dioden. Om snelle schakeltijden in pn-dioden te bereiken wordt de levensduur τ verkort door inbouwen van recombinatie-generatie-centra.

Voor toepassingen in gelijkrichters is de Schottky-diode te verkiezen boven de gewone pn-diode, bij dezelfde stroom kan met een kleiner oppervlak worden volstaan, is de schakeltijd korter en geldt over een groot gebied dat

$$J = J_s\,(e^{qV/kT} - 1).$$

Bij gebruik in zonnecellen is het gewenst dat de diodestroom volgens de hier gegeven formules zo klein mogelijk is. Omdat de zonnecellen veelal zodanig worden gebruikt dat de diode tamelijk ver voorwaarts staat is J_{rg} te verwaarlozen. Het trage diffusiemechanisme is bij uitstek geschikt om de diodestroom laag te houden. In zonnecellen zijn dan ook pn-dioden te verkiezen boven Schottky-dioden. In CCD's wordt gebruikgemaakt van gesperde dioden waarvan de lekstroom zo laag mogelijk moet zijn om de CCD's bij lage frequenties te kunnen bedrijven. In het algemeen zijn hier dus pn-diodes te verkiezen boven Schottky-diodes.

Praktische dioden

Bij het maken van dioden treden de volgende problemen op:

– het doteren van de halfgeleider door donor- of acceptoratomen op roosterplaatsen in te bouwen,

– het maken van een zuiver metaal-halfgeleidercontact,

– het lateraal begrenzen van de diode,

– het aanbrengen van een goed geleidend contact op de niet-diodekant van de halfgeleider.

Het doteren van halfgeleiders en in het bijzonder van silicium behoort tot de relatief gemakkelijkst oplosbare problemen.

Voor het maken van een zuiver metaal-halfgeleidercontact zijn de mogelijkheden beperkt. Bij het opbrengen van een metaal door sputteren of opdampen onder vacuüm is het moeilijk een schoon en oxydevrij halfgeleideroppervlak te verkrijgen. Schoonsputteren van het oppervlak leidt vaak tot beschadiging van het kristalrooster. Het neerslaan van metalen uit vluchtige verbindingen voorafgegaan door etsen in de gasfase lijkt de meest aantrekkelijke methode maar wordt weinig gebruikt. Meestal worden Schottky-dioden gemaakt zodanig dat het originele onzuivere halfgeleideroppervlak binnen het metaal komt te liggen. Twee soorten metalen komen hierbij in aanmerking. Metalen die SiO_2 kunnen reduceren zoals Al en Ti kunnen worden gebruikt om de SiO_2-laag op Si te doen verdwijnen. Metalen die silicides vormen zoals Pt kunnen worden gebruikt om een gedeelte van het silicium in silicide om te zetten. Bij beide groepen metalen zijn temperaturen tussen 500 en 1000 °C nodig zodat diffusie van het metaal in het silicium, afscheidingen bij rekristallisatie en dergelijke niet altijd te voorkomen zijn. Ook omdat het grensvlak hier niet een ideaal vlak wordt is het met deze methoden moeilijk perfecte Schottky-dioden te maken.

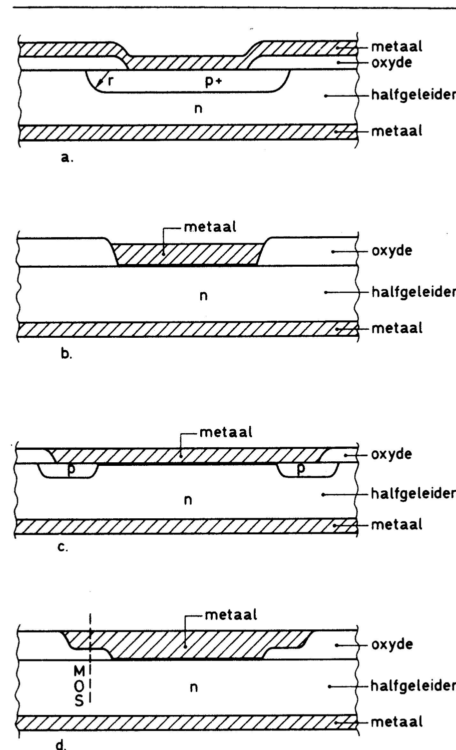


Fig. 3. Verschillende dioden met kromtestraaleffect. a. pn-diode met r > 0. b. Schottky-diode met r = 0. c. Schottky-diode omringd met pn-diode. d. Schottky-diode omringd met MOS-capaciteit.

Het lateraal begrenzen van de dioden is nodig vanwege het kromtestraaleffect wat maakt dat het elektrische veld aan de rand van de diode hoger is dan midden onder de diode. Pn-dioden kunnen gemakkelijk met een grote kromtestraal worden gemaakt en dotering en geometrie van losse dioden kunnen zo worden gekozen dat het veld aan de rand voldoende klein blijft om hier geen doorslag te veroorzaken. Bij Schottky-dioden kan de laterale begrenzing worden verkregen door omringing met een MOS-capaciteit of met een pn-diode. Bij zeer kleine Schottky-dioden, zoals in geïntegreerde schakelingen, is omringing met een MOS-capaciteit het gemakkelijkst. Voorbeelden van dioden zijn in figuur 3 schematisch aangegeven.

Slotopmerkingen

Om met succes dioden te kunnen gebruiken is kennis van de eenvoudige theorie aanbevelingswaardig. Naast theoretische overwegingen moet ook rekening worden gehouden met het gemak van fabricage. Dit laatste heeft gemaakt dat er momenteel meer pn-dioden worden vervaardigd dan Schottky-dioden, hoewel de laatste

honderd jaar ouder zijn en in theorie soms zijn te verkiezen, vooral als het om gelijkrichters gaat.

Referenties
1. E.S. Yang: *Fundamentals of semiconductor devices*. McGraw-Hill Book Company, New York, 1978.
2. S.M. Sze: *Physics of semiconductor devices*. Wiley-Interscience, New York, 1969.
3. E.H. Rhoderick, in Proc. 5th European Solid State Device Res. Conf., Société Française de Physique, Paris, 1975.
4. E.H. Rhoderick: *Metal-semiconductor contacts*. Clarendon Press, Oxford University Press, 1978.
5. A.S. Grove: *Physics and Technology of Semiconductor Devices*. John Wiley & Sons, Inc., New York, 1967.
6. 2nd E.C. Photovoltaic Solar Energy Conference at Berlin (West), Germany, D. Reidel Publishing Company, Dordrecht, Holland.

SiO_2-lagen op Si

M.W. Hillen en J.F. Verwey

De kwaliteit van geïntegreerde schakelingen wordt voor een belangrijk deel bepaald door de eigenschappen van de SiO_2-laag en van het Si/SiO_2-grensvlak. Daarom vormen SiO_2-lagen op Si een belangrijk onderzoekgebied van de fysica van halfgeleiderelementen.
Het belang van SiO_2-lagen en van de kwaliteit van deze lagen in moderne geïntegreerde schakelingen wordt aangegeven. De MOS (metaal-oxyde-silicium)-condensator blijkt bij uitstek een meetinstrument voor de eigenschappen van de SiO_2-laag. Beweeglijke ionen veroorzaken tal van ongewenste effecten in SiO_2-lagen. Op dit gebied is interessant onderzoek te doen aan de wijze waarop de ionen zich in de SiO_2-laag (al of niet in een chlooromgeving gegroeid) voortbewegen en hoe zij worden ingevangen aan het Si/SiO_2-grensvlak.

Het belang van SiO_2-lagen in geïntegreerde schakelingen

De grote vlucht die de elektronica heeft genomen, is te danken aan de ontwikkeling van de *monolitische geïntegreerde schakeling*. Hierbij zijn de elementen van een circuit in één substraat aangebracht. In de meeste gevallen is dit substraat een schijf monokristallijn silicium en vindt de fabricage plaats met de *planaire technologie*, waarin siliciumdioxyde (SiO_2) een belangrijke rol speelt.
In de planaire technologie wordt de siliciumschijf (de 'plak') voorzien van een laag SiO_2 (dikte ca. 100 nm), meestal door oxydatie bij hoge temperatuur (1000-1200 °C) in een zuurstofomgeving. In de oxydelaag worden vervolgens met een foto-lithografische techniek vensters geëtst, waardoor doopstoffen in het Si-kristal kunnen worden gebracht. Op deze wijze ontstaan dan n- of p-type gedoteerde gebieden. Dit proces wordt eventueel herhaald en tenslotte wordt een metaalpatroon over het oppervlak aangebracht dat de verbindingen vormt tussen de verschillende circuitelementen.

Voor de hier beschreven *maskerende werking* van het oxyde worden niet zo hoge eisen aan de kwaliteit van het SiO_2 gesteld. Om andere redenen gebeurt dat wel.

Allereerst is het voor een goede werking van transistors van belang dat de ruimteladingsgebieden van de pn-juncties niet aan het reactieve siliciumoppervlak komen, maar door een laag SiO_2 worden beschermd (*stabiliserende werking*). In de tweede plaats is het belangrijk dat er *weinig oppervlaktetoestanden zijn (passiverende werking)*. Oppervlaktetoestanden zijn plaatsen aan het Si/SiO_2-grensvlak, waar vrije ladingsdragers uit het Si ingevangen kunnen worden. Dat wil zeggen, zij introduceren energieniveaus in de verboden band van het silicium en worden daarom gemeten als functie van de energie ($cm^{-2}eV^{-1}$). Hun aantal is zeer afhankelijk van de wijze van oxyderen en van de nabehandeling.

Vooral de MOSFET (Metal-Oxide-Semiconductor Field-Effect Transistor) is hier zeer gevoelig voor. In figuur 1 is een p-kanaal-MOSFET getekend; het kanaal ontstaat doordat een voldoend grote negatieve spanning op de veldelektrode ('gate') gaten naar het oppervlak trekt. Er is nu een geleidende verbinding tussen het source- en het draingebied. Oppervlaktetoestanden kunnen het stroomtransport in het kanaal van een MOSFET ernstig belemmeren. Dit probleem was in het verleden zo groot, dat, hoewel in principe de werking van de veldeffect-transistor in 1930 bekend was, het toch tot de jaren zestig heeft geduurd voordat deze kon worden gefabriceerd.

Hoe het aantal oppervlaktetoestanden in de loop der jaren drastisch is verminderd, is te zien in figuur 2. Deze afname is voornamelijk bereikt door schonere fabricagemethoden en betere materialen. Voor een MOSFET in een modern geïntegreerd circuit is het aantal bewegende ladingsdragers in het kanaal ongeveer 10^{12} per cm^2. Het aantal oppervlaktetoestanden mag maar een fractie hiervan zijn om een ongestoord transport mogelijk te maken. Het minimaal te bereiken niveau is op dit moment ongeveer 10^8 cm^{-2}. Dit betekent dat slechts 1 op de 10^7 oppervlakte-atomen werkzaam is als oppervlaktetoestand. Voor de MOS-veldeffecttransistor is het bovendien van belang dat het oxyde onder de veldelektrode *weinig defecten* (minder dan 10 per cm^2) en *weinig lading* bevat. Defecten (zwakke plekken in het

Fig. 1. Schematische doorsnede van een p-kanaal-MOSFET. Enkele typische afmetingen zijn aangegeven. Let wel, de figuur is niet op schaal getekend. De metaallaag is zwart weergegeven.

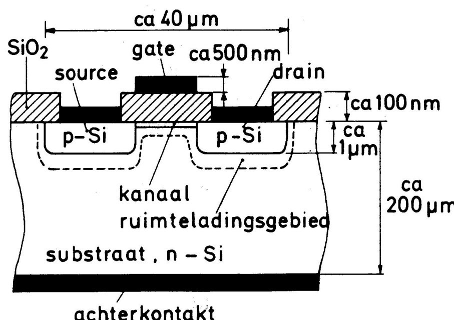
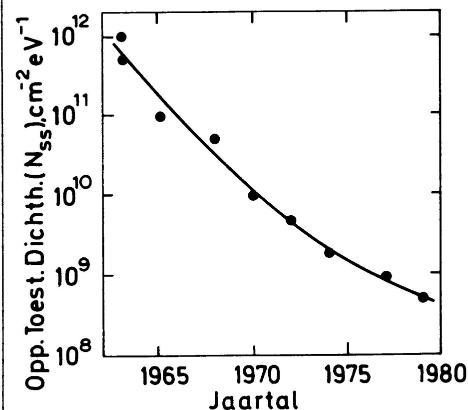
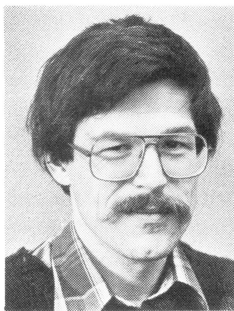


Fig. 2. Afname van de oppervlaktetoestandsdichtheid N_{ss} in het midden van de verboden band als functie van het jaartal. De gegevens zijn afkomstig van Goetzberger, e.a.[1]) en uit recente publicaties.



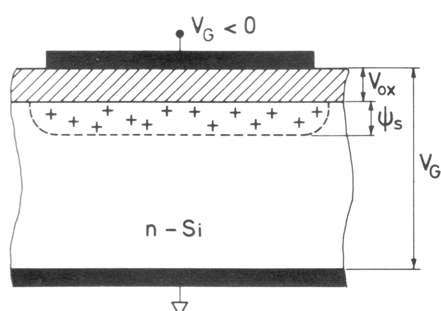


Ir. M.W. Hillen (31) studeerde technische natuurkunde aan de Rijksuniversiteit te Groningen. In 1976 behaalde hij het ingenieursdiploma na een onderzoek naar recombinatiestromen in epitaxiale diodes. Bij de onderzoekgroep Fysica van Halfgeleiderelementen verricht hij sindsdien onderzoek naar de invloed van beweeglijke lading in SiO_2 op de basisstroom van bipolaire transistoren, in het kader van een promotieopdracht onder leiding van prof. dr. J.F. Verwey.

oxyde) veroorzaken doorslag bij kleine elektrodespanningen en aangezien één defecte transistor in principe een gehele geïntegreerde schakeling (met honderden elementen) onbruikbaar maakt, is het onderzoek naar defectvrije oxydes in industriële sfeer van groot belang voor de opbrengst.

Lading in het oxyde kan beweeglijk zijn en vast. *Vaste oxydelading* bestaat uit positieve ladingen (bijvoorbeeld extra Si), die zich in de eerste 20 nm van het oxyde (gerekend vanaf het Si/SiO_2-grensvlak) bevinden. Zij kunnen geen lading met het silicium uitwisselen, zijn onbeweeglijk in een sterk elektrisch veld en onvermijdelijk aanwezig. Door een geschikte nabehandeling is hun aantal gering genoeg te maken. Tengevolge van deze ladingen is bijvoorbeeld in een n-kanaal-MOSFET het kanaal al aanwezig zonder dat een elektrodespanning is aangelegd. *Beweeglijke lading* in het oxyde blijkt ook positief te zijn en bestaat uit ionen van verschillende verontreinigingen die tijdens het technologisch proces worden ingebouwd. Deze ionen bewegen onder invloed van een elektrisch veld en veroorzaken zo instabiliteiten in de werking van een transistor. Zij zijn slechts te vermijden door de oxydatie onder zeer schone omstandigheden uit te voeren. Wij komen op de beweeglijke lading nog uitvoerig terug.

Fig. 3. Doorsnede van een p-kanaal-MOS-condensator. De elektrodespanning is altijd gedefinieerd ten opzichte van het substraat. Een negatieve elektrodespanning veroorzaakt een positieve ruimtelading, omdat de vrije elektronen van het grensvlak worden weggeduwd en er dan geïoniseerde vaste donoratomen in het kristal achterblijven.



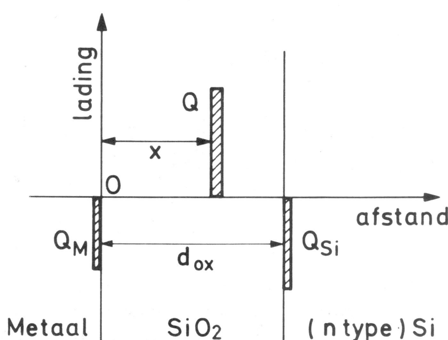
Onderzoek met behulp van de MOS-condensator

Uit het vorige is duidelijk geworden dat de MOS-veldeffecttransistor zeer gevoelig is voor de kwaliteit van het SiO_2 en van het Si/SiO_2-grensvlak. Een eenvoudiger structuur waarmee de eigenschappen van de SiO_2-laag en het Si/SiO_2-grensvlak kunnen worden bepaald, is de *MOS-condensator* (figuur 3). De spanning (V_G) die op de veldelektrode van de MOS-condensator wordt aangelegd verdeelt zich over het oxyde (V_{ox}) en de ruimteladingslaag in het silicium onder de elektrode (de oppervlaktepotentiaal ψ_s): $V_G = V_{ox} + \psi_s$. (De spanningsval over de rest van het silicium is verwaarloosbaar.)

Er is één elektrodespanning waarbij $\psi_s = 0$ (de energiebanden in het silicium buigen dan niet) en dat is de *vlakkebandenspanning* V_{FB}. Voor een ideale MOS-condensator is $V_{FB} = 0$, maar tengevolge van enerzijds het verschil in uittreepotentiaal tussen metaal en silicium en anderzijds oppervlaktetoestanden, vaste en beweeglijke lading in het oxyde, zal V_{FB} in de praktijk ongelijk aan nul zijn.

De invloed van een lading Q ter plekke x in het oxyde (figuur 4) op de vlakkebandenspanning is: $V_{FB} = -xQ/d_{ox}\,C_o$, waarin C_o de capaciteit van de oxydelaag is en d_{ox} de dikte. Voor een bewegende lading Q geldt dan dat deze aan het Si/SiO_2-grensvlak ($x = d_{ox}$) een maximale invloed uitoefent: $V_{FB} = -Q/C_o$ en aan het metaal/SiO_2-grensvlak totaal geen invloed uitoefent ($V_{FB} = 0$). Duidelijk is dan ook dat vaste lading aan het grensvlak en lading in oppervlaktetoestanden een bijdrage aan V_{FB} leveren.

Fig. 4. Een lading Q ter plekke x ($o \leqslant x \leqslant d_{ox}$) in het oxyde induceert beeldladingen in de metaallaag (Q_M) en in het Si (Q_{Si}). Uit Q_M kan V_{FB} worden berekend.





Prof. dr. J.F. Verwey (45) studeerde vaste-stofchemie aan de Rijksuniversiteit te Utrecht. Hij promoveerde daar in 1967 op een proefschrift, getiteld 'Photochemical Decomposition of Lead Halides'. Sindsdien is hij werkzaam op het Philips Natuurkundig Laboratorium, waar hij onder andere onderzoek heeft verricht op het gebied van betrouwbaarheid van geïntegreerde schakelingen en transistoren voor niet vluchtige informatieopslag. Sedert 1976 is hij als buitengewoon hoogleraar in de Technische Natuurkunde verbonden aan de Rijksuniversiteit te Groningen, waar hij de leiding heeft over de onderzoekgroep Fysica van Halfgeleiderelementen. Hij is voorzitter van de NNV-sectie Technische Halfgeleiderfysica en -elektronica en tevens voorzitter van de gelijknamige Contactgroep, die bij de Stichting voor Fundamenteel Onderzoek der Materie een aanvraag in voorbereiding heeft, als Werkgemeenschap binnen FOM te mogen gaan functioneren.

Het is van belang de hoeveelheid lading te kennen. De hoeveelheid beweeglijke lading is op de volgende manier te meten. De MOS-condensator wordt opgewarmd tot een temperatuur T_s (bijvoorbeeld 200°C). Dan wordt er een positieve spanning op de veldelektrode aangesloten $V_G = V_s > 0$. De beweeglijke ionen zullen nu onder invloed van het elektrisch veld in het oxyde naar het Si/SiO_2-grensvlak gaan. Het blijkt experimenteel dat de ionen *in* het oxyde blijven; zowel het Si/SiO_2- als het SiO_2/metaal-grensvlak laten geen ionen door.
Na een tijd t_s wordt de plak weer afgekoeld, waarbij de spanning V_s gehandhaafd blijft om de ionen op hun plaats te houden. Bij kamertemperatuur zijn de ionen minder beweeglijk en hier bepalen we de vlakkebandenspanning V_{FB}^+. Dan wordt dit experiment herhaald met gebruikmaking van een negatieve spanning $V_G = -V_s$, waardoor de ionen naar het metaal/SiO_2-grensvlak worden getrokken. Na afkoeling meten we een vlakkebandenspanning V_{FB}^- en uit de verschuiving $\Delta V_{FB} = |V_{FB}^+ - V_{FB}^-|$ volgt het aantal

beweeglijke ionen:

$$N_i = \frac{1}{q} \cdot C_o \cdot \Delta V_{FB}$$

Dit *BTS ('Bias-Temperature Stress')-experiment* is een standaardmeting in vele laboratoria.

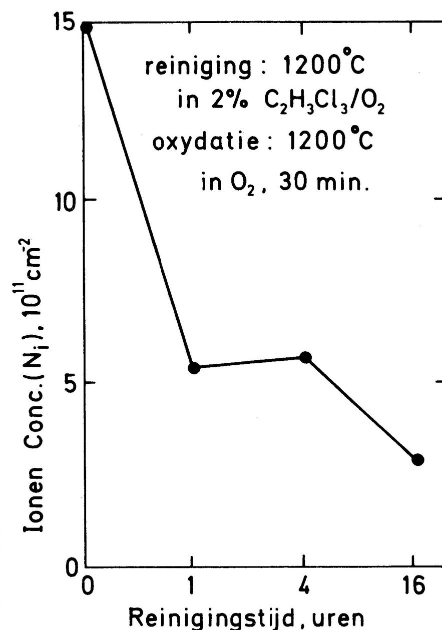
Ongewenste effecten veroorzaakt door beweeglijke ionen in SiO_2

Instabiliteit van de drempelspanning in een MOSFET

De drempelspanning is die spanning die op de veldelektrode van een MOSFET moet worden aangelegd, om juist een kanaal te laten ontstaan. Aangezien de drempelspanning afhankelijk is van de vlakkebandenspanning, kan deze variëren als er beweeglijke ionen in het oxyde aanwezig zijn. De eis is nu voor een moderne MOSFET: een spanningsverschuiving ten gevolge van instabiliteiten van maximaal 0,05 V. Dit komt neer op minder dan 2×10^{10} beweeglijke ionen per cm^2 (minder dan 1 op de 100.000 atomen in een monolaag).

In de technologie zijn verschillende methoden in gebruik om tot zo'n lage concentratie te komen. We noemen hier slechts een zeer bekende en misschien wel de meest effectieve: schoon werken. Eén aspect hiervan is bijvoorbeeld het schoonmaken van de ovenbuis vóór de oxydatie. Dit schoonmaken gebeurt door gedurende enkele uren HCl door de ovenbuis te leiden. Hierdoor wordt de ovenwand uitgeloogd en gedurende enige tijd daarna verwachten we geen verontreinigingen uit de ovenwand. Dat dit effect inderdaad optreedt blijkt uit figuur 5, waar de concentratie van beweeglijke ionen in het oxyde is weergegeven als functie van de tijd dat de oven werd schoongemaakt vóór de oxydatie.

Fig. 5. De invloed van de tijd, dat de oxydatie-oven wordt schoongespoeld op de concentratie van beweeglijke ionen in het oxyde (N_i) verkregen door oxydatie na het schoonspoelen. De chloorverbinding, waarmee wordt schoongespoeld is $C_2H_3Cl_3$, dat bij hoge temperatuur equivalent is met het veel agressievere HCl.



Creatie van extra oppervlaktetoestanden door beweeglijke ionen

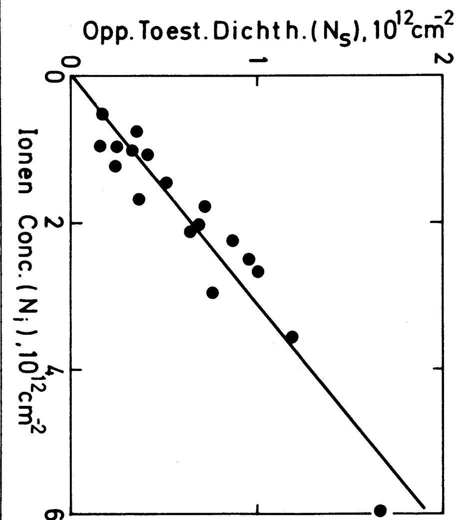
Al in 1968 veronderstelde Goetzberger[1] dat een positief ion aan het Si/SiO_2-grensvlak een elektron in het Si kan aantrekken. Dit elektron kan dan niet meer vrij door het Si-kristal bewegen en er is een oppervlaktetoestand ontstaan. Voor de beweeglijke ionen betekent dat, dat er meer oppervlaktetoestanden zouden moeten zijn als de ionen zich aan het Si/SiO_2-grensvlak bevinden dan als de ionen aan het metaal/SiO_2-grensvlak zijn.

Dat een dergelijk effect inderdaad optreedt is pas zeer onlangs aangetoond. Het probleem hierbij is dat het aantal oppervlaktetoestanden moet worden bepaald bij een temperatuur, waarbij de ionen niet meer beweeglijk zijn ($T <$ -30 °C). Snel[2] heeft een lineair verband gevonden tussen het aantal beweeglijke ionen aan het Si/SiO_2-grensvlak en het aantal oppervlaktetoestanden in de middelste 0,8 eV van de verboden band van Si (met een methode die werkt bij $T =$ -196 °C). Het is echter niet zo dat elk ion juist één oppervlaktetoestand veroorzaakt (figuur 6).

Veroudering

Geen enkel halfgeleiderelement heeft het eeuwige leven, ook al wordt het geheel volgens de voorschriften in een schakeling gebruikt. De *levensduur* van een MOSFET wordt vaak bepaald door de tijd die het duurt voor een gate-oxyde elektrisch doorslaat. Aangezien de levensduur jaren kan en moet bedragen, worden levensduurtests vaak onder verzwaarde omstandigheden (hoge temperatuur en hoge elektrische spanning) uitgevoerd. Het BTS-experiment is hier zeer geschikt voor. *Veroudering* is het verschijnsel dat de kwaliteit van een SiO_2-laag afneemt bij langdurig normaal gebruik. Het uit zich in een fatale elektrische doorslag van het oxyde. Veroudering kan worden onderzocht met BTS-experimenten, die worden uitgevoerd bij een spanning, die lager is dan de doorslagspanning. De doorslagspanning is de spanning waarbij het oxyde direct doorslaat, dus vóórdat het BTS-experiment wordt toegepast. Figuur 7 toont experimentele resultaten, waaruit blijkt dat als de concentratie van beweeglijke ionen in het oxyde hoger is, de veroudering sneller optreedt.

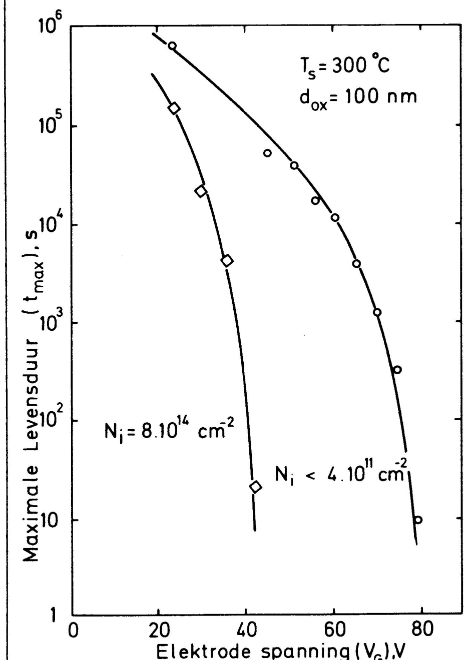
Fig. 6. Verband tussen de oppervlaktetoestandsdichtheid en het aantal beweeglijke ionen N_i. N_i is bepaald met de TSIC-methode (zie verderop).



Een mogelijke verklaring is dat bij positieve elektrodespanning het gecombineerde effect van positieve elektrodelading en positieve ionlading aan het Si/SiO_2-grensvlak elektronen in het oxyde trekt. Deze elektronen, versneld door het veld, veroorzaken stoot-ionisatie, waardoor weer extra elektronen ontstaan, enzovoort. Dit lawine-effect veroorzaakt dan doorslag van de SiO_2-laag.

Of dit inderdaad het verouderingsproces is, is echter nog niet duidelijk. Dat er

Fig. 7. Maximale levensduur t_{max} van een MOS-condensator tijdens een verouderingsexperiment als functie van de aangelegde spanning V_s tijdens het BTS-experiment. Parameter is de concentratie van beweeglijke ionen N_i [2].



ionen bij betrokken zijn lijkt waarschijnlijk uit de experimenteel gevonden afhankelijkheid van de veroudering van de toegepaste elektrodespanning en de gebruikte temperatuur. Toenemende kans op veroudering blijkt op te treden bij het toepassen van steeds dunnere oxyden en bij het gebruik van elektronen- en laserbundels in de technologie. Dit maakt verder onderzoek van de verouderingsprocessen noodzakelijk.

De eigenschappen van beweeglijke ionen in SiO_2-lagen

Uiteraard kunnen de ongewenste effecten van beweeglijke ionen worden onderdrukt, door te zorgen dat zij niet in de oxydelaag worden ingebouwd. Er is dan ook veel onderzoek gedaan naar fabricage van halfgeleiderelementen in een zeer zuivere omgeving, maar enige verontreiniging is vaak niet te vermijden, gelet op het gewenste lage niveau van 1 beweeglijk ion op de 100.000 oppervlakte-atomen. Onderzoek naar de aard en het gedrag van de ionen kan ons in dit verband helpen om de ongewenste effecten te onderdrukken.
Uit de hierboven beschreven BTS-experimenten bleek niet alleen dat ionen onder invloed van een elektrisch veld door het oxyde bewegen maar ook dat een bepaalde hoeveelheid energie moet worden toegevoerd (elektrisch of thermisch) om ze van de grensvlakken vrij te krijgen. Kennelijk zijn er aan de grensvlakken plaatsen waar de ionen een minimale energie bezitten (*invangcentra*).
Deze beide eigenschappen kunnen met afzonderlijke experimenten nader worden onderzocht. Overigens merken we op dat het experimenteel blijkt dat de ionen *in* het oxyde blijven; zowel het Si/SiO_2- als het SiO_2/metaal-grensvlak laten geen ionen door. Omdat we in de eerste plaats geinteresseerd zijn in de eigenschappen van het Si/SiO_2-grensvlak, beschouwen we in het volgende de beweging van ionen vanaf het Si/SiO_2-grensvlak naar het SiO_2/metaal-grensvlak toe (dus onder invloed van een negatieve elektrodespanning).

De invangcentra aan het Si/SiO_2-grensvlak en de aard van de ionen

Een methode om de energie te bepalen waarmee de ionen aan een grensvlak ingevangen zitten is de *thermisch gestimuleerde ionenstroom (TSIC)-methode*. Hierbij wordt na een BTS-experiment onder positieve elektrodespanning (alle ionen bevinden zich hierna aan het Si/SiO_2-grensvlak) de temperatuur van de MOS-condensator langzaam verhoogd, van 0 tot 350 °C, met een constante opwarmsnelheid (0,1-1 °C/s). Er wordt een zo grote negatieve spanning op de elektrode gezet, dat elk ion dat uit het invangcentrum weet te ontsnappen direct naar de metaalelektrode wordt getrokken. De kans om te ontsnappen is sterk temperatuurafhankelijk en neemt toe met toenemende temperatuur. Aangezien voor elk ion dat het oxyde oversteekt, zich één elektron in het uitwendige circuit verplaatst, meten we de laadstroom van de condensator. Een TSIC-meting is getoond in figuur 8. In het bewuste temperatuurgebied zien we twee stroompieken. Dat de stroom weer nul wordt betekent dat alle ionen zijn overgestoken.

Wat de *aard* van de mobiele ionen betreft: algemeen wordt aangenomen dat Na^+-ionen de voornaamste oorzaak van instabiliteiten zijn. Na^+-ionen kunnen vrijkomen tijdens de fabricage (bijvoorbeeld uit de wand van de oxydatieoven of uit het schuitje vanwaaruit het metaal wordt opgedampt). Het blijkt dat de piek bij lage temperatuur in de TSIC-meting door Na^+ wordt veroorzaakt. Het ligt voor de hand voor de tweede piek aan een groter ion te denken (heeft meer thermische energie nodig om uit het invangcentrum te worden geëmitteerd). Zeer waarschijnlijk hebben we hier te maken met K^+-ionen, zoals proeven met opzettelijk ingebracht K^+ (uit een K^+ bevattende oplossing of door K^+-implantatie) aantoonden[3]). Deze mening wordt echter niet door iedereen gedeeld. Het probleem is dat bij het opzettelijk aanbrengen van de ene verontreiniging, de andere vaak moeilijk is te vermijden.

We kunnen ons de ionen in de invangcentra als volgt (figuur 9a) voorstellen: n_t ionen zitten in de potentiaalkuil aan het grensvlak gevangen. Er is een bepaalde kans dat zij een energie E_o opnemen, waarmee ze uit de potentiaalkuil kunnen ontsnappen, waarna ze door het elektrisch veld direct worden afgevoerd. Deze kans $P(E)$ wordt beschreven met een Boltzmannfactor. Het aantal ontsnapte ionen per tijdeenheid hangt af van deze kans en van het aantal aanwezige ionen:

$$\frac{dn_t}{dt} = -n_t \, . \, P(E) = -n_t \, . \, s \exp\left(-\frac{E_o}{kT}\right).$$

s heeft hier de dimensie van frequentie en wordt geassocieerd met de atomaire trillingsfrequentie (orde 10^{12} Hz). Dit model blijkt de TSIC-curves heel aardig te verklaren. Een verfijning van het model bereiken we door aan te nemen dat niet alle ionen precies dezelfde ontsnappingsenergie E_o nodig hebben, maar dat daar een zekere spreiding ΔE in mogelijk is (figuur 9b). Het is mogelijk om uit de TSIC-curve de energieverdeling van de ionen in de invangcentra te bepalen.

De beweeglijkheid van de Na^+- en K^+-ionen

Bekeken we met de TSIC-methode het emissiegedrag, het is ook mogelijk om het transport van de ionen van de ene kant van het oxyde naar de andere kant te bekijken. Brengen we bijvoorbeeld na een BTS-experiment de MOS-condensator op een zo hoge temperatuur dat een groot aantal ionen voldoende energie heeft om uit de invangcentra te ontsnappen en variëren we *langzaam* de elektrodespanning, dan zullen de ionen het oxyde oversteken zodra het elektrisch veld van teken omkeert. Dit gebeurt in de *TVS('Triangular Voltage Sweep')-methode*. Voor de elektrodespanning wordt hier een langzaam verlopende driehoekspanning (helling 0,1-1 V/s) gebruikt.

Bepalen we nu met de TSIC-meting die temperatuur, waarbij de Na^+-ionen wel en de K^+-ionen nog niet genoeg energie hebben bereikt om te ontsnappen, dan kunnen we het transport van alleen de Na^+-ionen onderzoeken met een TVS-meting op deze temperatuur (225°C, zie figuur 8).

Om het transport van de ionen onder invloed van een elektrisch veld te kunnen beschrijven, moeten we allereerst hun beweeglijkheid μ kennen. Beschouwen we het oxyde als een aaneenschakeling van potentiaalputjes met gemiddelde diepte W (figuur 10), dan kan de beweeglijkheid worden geschreven als

Fig. 8. Het TSIC-experiment. De stroom ten gevolge van de beweging van de ionen I_i als functie van de temperatuur[3].

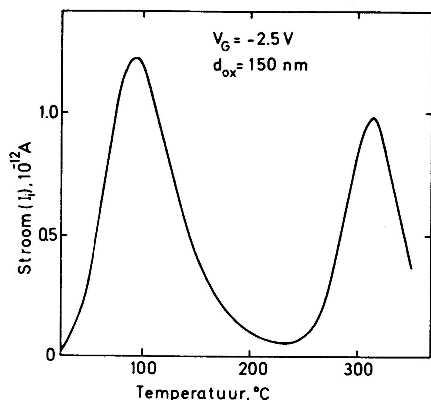
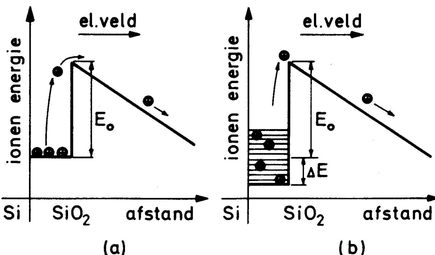


Fig. 9. Schematisch model van invangcentra voor ionen aan het Si/SiO_2-grensvlak. Als een ion is ontsnapt, wordt het direct door het elektrisch veld afgevoerd. (a) centra met één enkele energie E_o, (b) centra met spreiding in energie $E_o - \Delta E < E < E_o + \Delta E$.



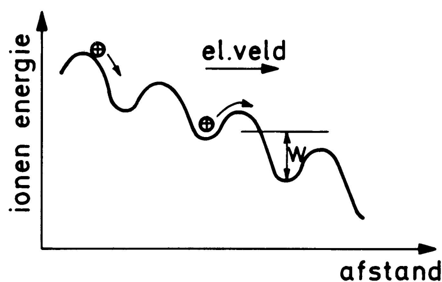


Fig. 10. Schematische voorstelling van de ionenbeweging in SiO_2, onder invloed van een elektrisch veld.

$\mu = (\mu_0/T) \exp[-W/kT]$.
Zo'n verband wordt experimenteel inderdaad gevonden. Het ion springt als het ware van het ene putje in het volgende. W is de activeringsenergie van de *drift* (beweging onder invloed van een elektrisch veld). De beweeglijkheid van Na^+- en K^+-ionen kan op twee manieren worden gemeten. In de eerste plaats door bij een bepaalde temperatuur de elektrodespanning *plotseling* om te keren en de oversteektijd te meten van die ionen die genoeg energie hadden om niet ingevangen te zijn. Het nadeel van deze methode is dat bij lage temperatuur dit aantal 'vrije' ionen erg klein kan zijn. In de tweede plaats met behulp van de TVS-methode, waarbij de elektrode-spanning *langzaam* wordt gevarieerd. Zo'n TVS-meting is in figuur 11 weergegeven. Door de hoge temperatuur hebben zowel de Na^+- als de K^+-ionen voldoende energie om uit de invangcentra te ontsnappen. De beweeglijkheid van Na^+-ionen is zo groot dat zij het oxyde direct oversteken als het elektrisch veld in het oxyde van teken omkeert (bij V_G = 0,1 V). In de TVS-meting veroorzaken zij een stroompiek bij die spanning. De K^+-ionen beginnen bij deze spanning ook aan hun overtocht. Hun beweeglijkheid is lager en het duurt enige tijd voor zij zijn overgestoken. Hun snelheid wordt wel steeds groter omdat de elektrodespanning en dus het elektrisch veld in het oxyde steeds toeneemt. Als de eerste K^+-ionen de overkant bereiken is de stroom maximaal. Uit de plaats van dit maximum als functie van temperatuur en helling van de langzaam verlopende driehoekspanning kan de beweeglijkheid worden bepaald[4]. Een overzicht van recente bepalingen van de beweeglijkheid van Na^+- en K^+-ionen volgens beide methoden is weergegeven in figuur 12. De resultaten sluiten goed bij elkaar aan.

Beweeglijke ionen in chlooroxyden
Een nieuwe ontwikkeling in de fabricage van MOSFET's is de toepassing van chloorverbindingen (HCl of 1.1.1-$C_2H_3Cl_3$), niet alleen om de ovenbuis te reinigen, maar ook als toevoeging tijdens de oxydatie. Mits dit niet in te grote hoeveelheden gebeurt, treden heilzame effecten op: het aantal defecten wordt minder (figuur 13) en de veroudering wordt minder. Of het aantal oppervlaktetoestanden groter of kleiner wordt hangt sterk af van de oxydatie-omstandigheden.

Het chloor heeft een zeer merkwaardige invloed op de beweeglijke ionen, speciaal op Na^+. Het blijkt dat het chloor in het

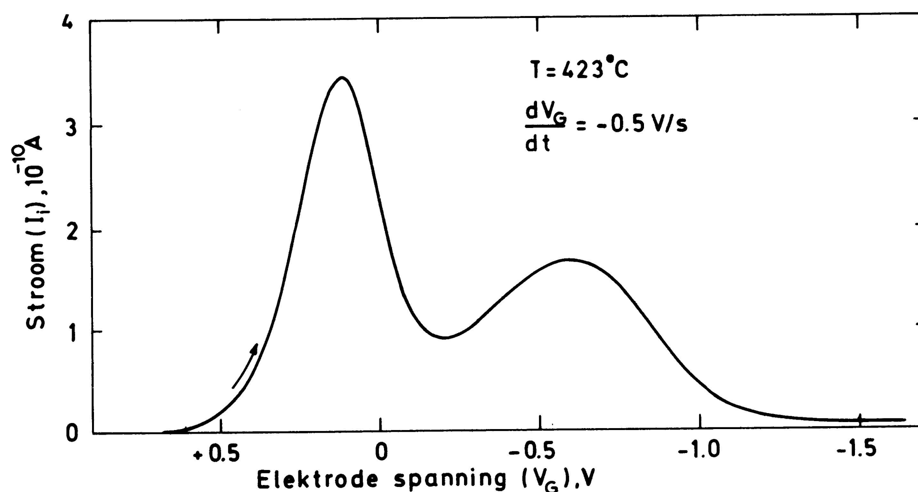
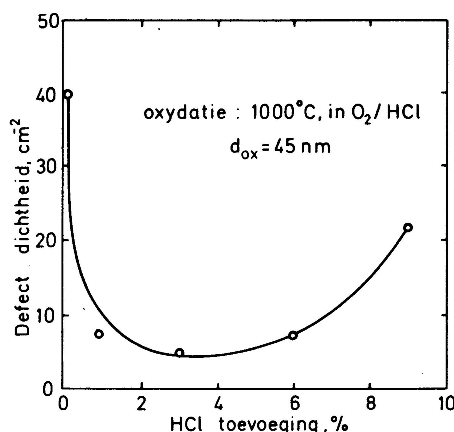
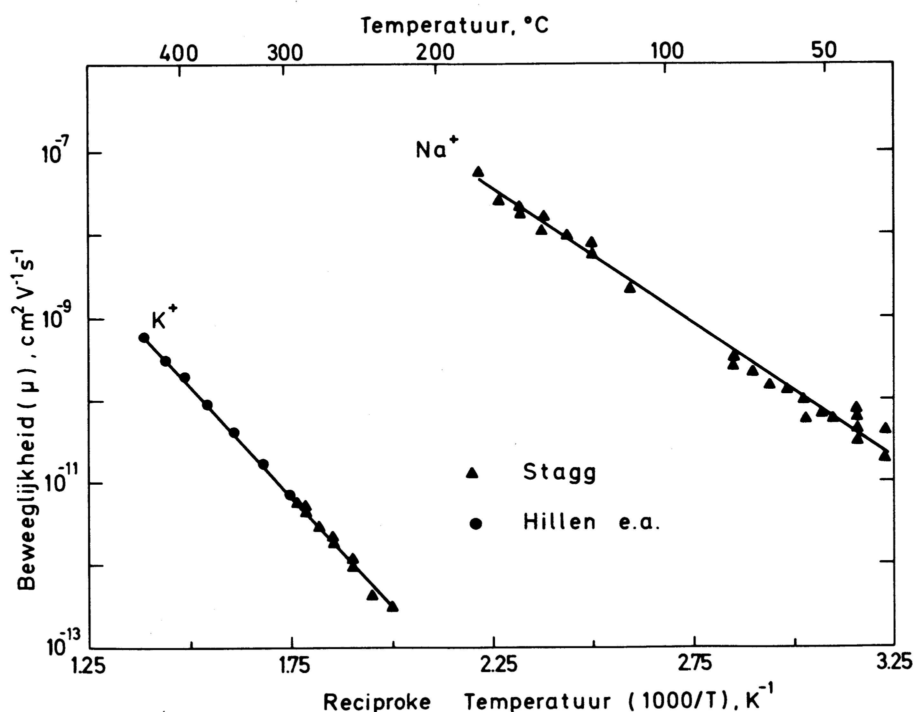


Fig. 11. Voorbeeld van een TVS-meting, waaruit de beweeglijkheid van K^+-ionen kan worden berekend[4]. De laadstroom ten gevolge van de capaciteit van het oxyde is onderdrukt.

Fig. 12. Beweeglijkheid van Na^+- en K^+-ionen als functie van de temperatuur. De metingen zijn afkomstig van Stagg[5] (oversteektijd-meting) en Hillen, e.a.[4]) (TVS-meting). De waarden voor Na^+ gelden zowel voor oxyden gegroeid in O_2 als voor chlooroxyden.

Fig. 13. Defectdichtheid in het oxyde als functie van de HCl-toevoeging tijdens de oxydatie[6].





oxyde wordt ingebouwd dicht bij het Si/SiO_2-grensvlak. Worden nu de Na^+-ionen in een TVS-experiment bij 225 °C naar het Si/SiO_2-grensvlak gedreven (met een positieve elektrodespanning) dan komen de ionen als de spanning van teken wordt omgedraaid niet meer terug! Bovendien blijkt uit een vlakkebanden-spannings-meting dat het effect dat Na^+-ionen aan het Si/SiO_2-grensvlak op de vlakkebandenspanning hebben geheel verdwenen is: de Na^+-ionen hebben kennelijk hun lading verloren, ze zijn *geneutraliseerd*. Dat dit effect niet definitief is, zien we aan TSIC-metingen. Deze metingen worden uitgevoerd na een zodanig BTS-experiment dat de Na^+-ionen naar het Si/SiO_2-grensvlak zijn gebracht. In figuur 14 worden TSIC-metingen getoond aan MOS-condensators met en zonder chloor in het oxyde. Beide oxydes zijn opzettelijk met Na^+-ionen verontreinigd.

In de TSIC-curve van de MOS-condensator zonder chloor in het oxyde, herkennen we de stroompiek ten gevolge van 'vrije' Na^+-ionen, die ook te zien was in figuur 8. Dat de plaats, vorm en grootte van deze Na^+-piek in beide figuren niet gelijk is, komt door verschillen in opwarming, veld in het oxyde en aantal Na^+-ionen. Van het stroompiekje bij 250 °C vermoeden we dat het K^+-ionen zijn, ingebracht tijdens de metallisatie. Hun aantal is echter veel kleiner dan het aantal Na^+-ionen.

In de TSIC-curve van de MOS-condensator met chlooroxyde ontbreekt de bijdrage van de 'vrije' Na^+-ionen. Pas bij ruim 300 °C komen de geneutraliseerde Na^+-ionen weer in ionvorm vrij en steken het oxyde over. Deze hoge temperatuur bekent dat de invangcentra diep zijn, dat wil zeggen: er is een grote energie E_0 nodig om de ionen uit de centra te laten ontsnappen.

De conclusie is, dat bij toevoeging van chloor aan oxyden de invangcentra voor Na^+-ionen aan het Si/SiO_2-grensvlak zo worden gewijzigd, dat het Na^+-ion dieper wordt ingevangen en bovendien zijn lading verliest. Het mechanisme van deze vorm van invangen is nog een onbegrepen zaak. Omdat het chloor aan het grensvlak wordt ingebouwd, is de beweeglijkheid van Na^+-ionen in het oxyde onafhankelijk van de toevoeging van chloor.
Dat er voor K^+-ionen in chlooroxyden ook merkwaardige effecten optreden, blijkt uit TVS-metingen. Wat er aan de hand is zal nog nader moeten worden onderzocht.

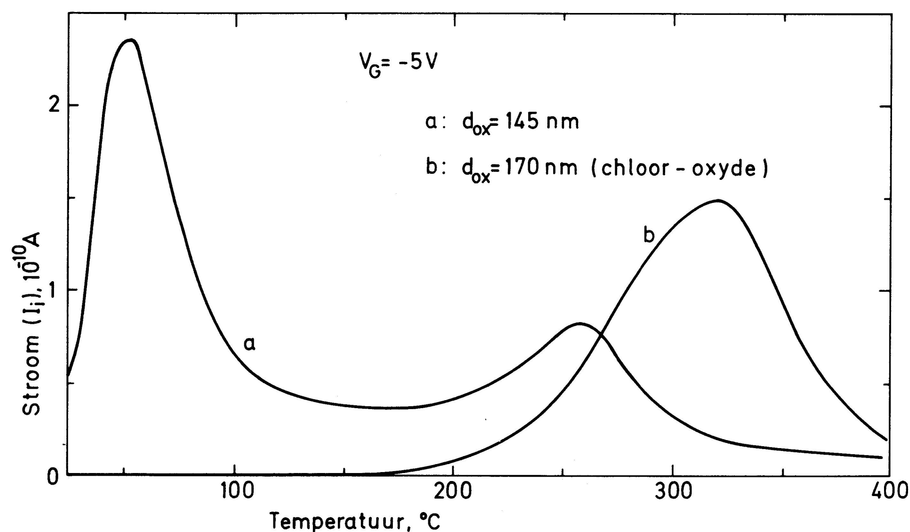


Fig. 14. TSIC-experiment aan MOS-condensators met oxyden gegroeid in O_2 (curve a) en in 3% $C_2H_3Cl_3/O_2$ (curve b). Na de oxydatie zijn de oxyden met Na^+-ionen uit een Na^+ bevattende oplossing verontreinigd. De ionen bewegen vanaf het Si/SiO_2-grensvlak naar het metaal/SiO_2-grensvlak toe.

Literatuur

1. A. Goetzberger, E. Klausmann, M.J. Schulz, C.R.C. Critical Reviews in Solid-State Sciences (1976) 1.
2. J.F. Verwey, Infos '79 conferentie, Durham; uit te geven in Inst. Phys. Conf. Series nr. 50 (1980).
3. M.W. Hillen: *The Physics of SiO_2 and its interfaces*. Pergamon 1978, p. 179-183.
4. M.W. Hillen, G. Greeuw, J.F. Verwey, J. Appl. Phys. **50** (1979) 4834.
5. J.P. Stagg, Appl. Phys. Lett. **31** (1977) 532.
6. C.M. Osburn, J. Electrochem. Soc. **121** (1974) 809.

Boekrecensies

W. Thirring,
Lehrbuch der Mathematischen Physik,
Band 1: Klassische Dynamische Systeme.
Springer-Verlag, Wenen, 1977.
XIII, 255 blz., 58 fig., DM 36,–, paper.

W. Thirring,
Lehrbuch der Mathematischen Physik,
Band 2: Klassische Feldtheorie.
Springer-Verlag, Wenen, 1978.
X, 258 blz., 70 fig., DM 36,–, paper.

V.I. Arnold,
Mathematical methods of classical mechanics.
Springer-Verlag, Berlin, Heidelberg, New York, 1978.
X, 462 blz., 246 fig., DM 54,–, gebonden.

Allereerst wil ik trachten de twee grondgedachten van de moderne differentiaalmeetkunde te isoleren. Daarmee heb ik dan tevens het gemeenschappelijk fundament van de thans te bespreken werken aangegeven.

Enerzijds wenst men het begrip *inhoud* (volumen) zodanig te formuleren dat het geldig is voor willekeurige n-dimensionale ruimten. Het is bekend dat in $\mathbb{R}^n$ de determinant functie $\det(v_1, \ldots, v_n)$ op het teken na het volumen geeft van het parallelepipidum dat opgespannen wordt door de vectoren $v_1, \ldots, v_n$. De determinantfunctie is lineair in elk van zijn argumenten, en bovendien is ze anti-symmetrisch. Men zegt dat de determinant een alternerende n-vorm is. Indien men nu iets analoog wil doen voor k vectoren in een n-dimensionale ruimte ($k \leqslant n$) dient men het begrip *alternerende k-vorm* te introduceren.

Maar behalve de notie van volumen wil men òok het begrip *oppervlak* generaliseren. Intuïtief is een cirkel een één-dimensionaal oppervlak in $\mathbb{R}^2$ en een bol een twee-dimensionaal oppervlak in $\mathbb{R}^3$. De vraag is natuurlijk: hoe parametriseren wij deze oppervlakken? Dit gebeurt middels het begrip *deelvariëteit*. De cirkel is een één-dimensionale deelvariëteit van $\mathbb{R}^2$, en in het algemeen spreekt men van een m-dimensionale deelvariëteit in $\mathbb{R}^n$ ($m \leqslant n$). Een deelvariëteit beschrijft men met behulp van 'kaarten', die dan in een 'atlas' worden verenigd. Een nette deelvariëteit zal men tevens willen voorzien van een rand en een oriëntatie. Indien zulks mogelijk is verkrijgt men een fraaie synthese van de begrippen alternerende k-vorm en (k + 1)-dimensionale deelvariëteit met oriëntatie en rand via de stelling van Stokes.

De lezer dient zich evenwel te realiseren dat de moderne differentiaalmeetkunde op het eerste gezicht nog maar in weinig gelijkt op wat een fysicus zich traditio-

neel had voorgesteld, en ze bovendien een zwaar en zeer formeel (zo niet abstract) begripsapparaat benodigt. Kortom, men dient zich te realiseren dat dit vak niet eenvoudig is, doch tevens dat het in zijn huidige vorm stormachtige ontwikkelingen doormaakt. Wil men het vak voor fysici tot leven laten komen dan is een goede, uitgebreide, en zorgvuldig *motiverende* inleiding van het allergrootste belang. Voor mathematici is zulk een inleiding gegeven door Spivak [1, 2], maar tot op heden ontbrak het aan literatuur [3] die duidelijk gericht was op fysische problemen. In deze leemte hebben nu Thirring en Arnold willen voorzien. Bij een geslaagde opzet zouden we hun daarvoor zeer erkentelijk moeten zijn. Maar we dienen wel te bedenken dat hun boeken bedoeld zijn als *leerboeken*, niet als wetenschappelijke monografieën.

De klassieke mechanica met haar 'nevenvoorwaarden' en corresponderende gegeneraliseerde coördinaten geeft op natuurlijke wijze aanleiding tot de introductie van deelvariëteiten, raakruimten, en meer van dat schoons dat dient tot opluistering van de bewegingsvergelijkingen. Ook het begrip inhoud komt op natuurlijke wijze naar voren via het theorema van Liouville. Een passend kader ter introductie van de nieuwe begrippen is hier dus duidelijk aanwezig.

Thirring's boek 'Klassische dynamische Systeme' bestaat uit drie gedeelten: een derde is gewijd aan het mathematische formalisme, een derde aan Hamilton systemen (*met* een bewijs van het KAM-theorema!) en niet-relativistische meerdeeltjes problemen, en het laatste deel aan de relativiteitstheorie. De eruditie van de schrijver is imponerend, de veelheid van problemen is fascinerend, maar de presentatie lijkt mij volstrekt ongeschikt om dit boek als leerboek te laten functioneren. We zullen ter staving van deze bewerking eens het eerste deel (het mathematische formalisme) bezien dat absoluut essentieel is voor al wat volgt. Niet alleen is de stijl een verschrikking (praktisch iedere zin is genummerd) maar ook ontbreekt elke motivatie. Het geheel is een dorre opeenvolging van definities, lemma's, stellingen en opmerkingen, en er wordt een algemeenheid en een abstractie nagestreefd die, me dunkt, zelfs Bourbaki tot schuchterheid en bescheidenheid zouden hebben gestemd. Ook de formulering is veel te kort en te strak gehouden (e.g. het lemma van Poincaré is in een vraagstuk ondergebracht). De stijl en opzet van het 'samenvattende' collegedictaat waarbij de collegegever een essentiële rol speelt, is overal nog speurbaar; te lichtvaardig is deze stijl in het boek gehandhaafd.

Ik heb mij afgevraagd hoe dit boek in de nederlandse context zou kunnen functioneren. Thirring zelf geeft een suggestie: hij werkt 'auf einführende Vorlesungen aufbauend'. Goed, ik heb eens een inleidend (prekandidaten) collegedictaat [4] ter hand genomen en nog eens te meer ontdekt wat aan Thirring's boek ontbreekt: leesbaarheid. Hoe dan ook, zelfs met een inleidend college à la [4] is, vrees ik, de nederlandse student nog onvoldoende toegerust om het mathematisch formalisme snel en doeltreffend te kunnen verwerken. Ik kan derhalve 'Klassische dynamische Systeme' alleen aan die docenten (en studenten) in overweging geven die èn zeer gecharmeerd zijn van een formele, 'mathematisch lückenlose' opbouw, èn over een zee van tijd beschikken, èn de steun hebben van een vakgroep Wiskunde die het vele, wiskundig benodigde grondwerk willen verrichten.

Thirring's tweede recente boek, 'Klassische Feldtheorie', is in bijna dezelfde stijl geschreven als het eerste deel. Na een inleidend hoofdstuk komen aan de orde het elektromagnetische veld van gegeven ladingsverdelingen, het veld in de aanwezigheid van een geleider, en tenslotte gravitatie. Beide delen bevatten vele opgaven *met* de bijbehorende oplossingen.

Alle vier delen van de door Thirring geplande reeks over Mathematische Fysica zullen in het engels worden vertaald; het eerste deel is reeds verschenen, eveneens bij Springer (nu Berlin, etc.). Voor slechts weinige guldens meer krijgt men dan een gebonden boek, een betere typografie en een iets betere formulering. Ik betreur echter dat hierdoor deze boeken worden gefixeerd in een vorm die nog te zeer tijdsgebonden en daardoor nog te weinig definitief is. Maar wellicht is dat een kenmerk van baanbrekend werk.

V.I. Arnold's boek, 'Mathematical methods of classical mechanics', ademt een geheel andere sfeer. Ook hier moderne differentiaalmeetkunde, maar wat een verschil tussen de mathematicus Arnold en de fysicus Thirring! Arnold heeft zijn boek zeer zorgvuldig geschreven [5, 7]. Hij illustreert zijn definities en beweringen uitgebreid aan de hand van voorbeelden en figuren en tracht een en ander zoveel mogelijk te visualiseren. Het boek is met grote zorg opgebouwd (het heeft een 'gelaagde' structuur) zó dat de lezer steeds goed gemotiveerd blijft. En een stelling wordt vaak niet in haar abstractste en algemeenste formulering bewezen maar juist zo (bijv. door een eenvoudig geval te bekijken) dat de bewering begrijpelijk en doorzichtig wordt. Toegegeven, ook hier wordt af en toe enige zelfwerkzaamheid verlangd, maar in de praktijk is reeds gebleken dat er met dit boek uitstekend te werken valt. Het laat zich ook voortreffelijk als aanvulling op Goldstein [6] gebruiken: begrippen als Legendretransformatie, canonieke transformatie, integraalinvariant van Poincaré-Cartan, of 'the principle of least action' vindt men bij Arnold op heldere en, zo men wil, alternatieve wijze behandeld. Dit boek bestaat eveneens uit drie gedeelten: de mechanica volgens Newton, volgens Lagrange (introductie van deelvariëteiten), en volgens Hamilton (introductie van alternerende vormen, etc.). In een uitgebreide reeks appendices worden tenslotte geavanceerde onderwerpen behandeld. Ik zou (theoretisch) fysici dit boek ten zeerste willen aanbevelen.

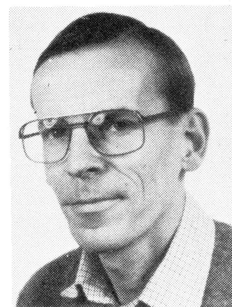
Referenties

1. M. Spivak, 'Calculus on manifolds', W.A. Benjamin (1965). De hoofdstukken 4 en 5 vormen een korte inleiding.
2. M. Spivak, 'A comprehensive introduction to differential geometry', 2nd Edition, Publish or Perish, Inc., P.O. Box 7108, Berkeley, CA 94707 (1979). Met name vol. I; er zijn nog vier delen.
3. Een vroege uitzondering is: Res Jost, 'Poisson Brackets (An unpedagogical lecture)', Rev. Mod. Phys. **36** (1964) 572-9.
Nog iets vroeger, en zeker leesbaar is: H. Flanders, 'Differential forms, with applications to the physical sciences', Academic Press (1963). Het aantal fysische toepassingen in dit boek is echter beperkt.
4. F. Takens, 'Differentiaalrekening in $\mathbb{R}^n$', Mathematisch Instituut, R.U. Groningen.
5. Zie ook: V.I. Arnold, 'Ordinary differential equations', The MIT Press, Cambridge (Mass.) and London (1973). Van dit boek bestaat nu tevens een goedkope paperback editie.
6. H. Goldstein, 'Classical Mechanics', Addison-Wesley (1950).
7. Voor diegenen die aan frans de voorkeur geven boven engels: van beide boeken van Arnold bestaat ook een (goedkope) franse editie.

J.L. van Hemmen

Beeldopnemer met weerstandselektroden

M.G. Collet

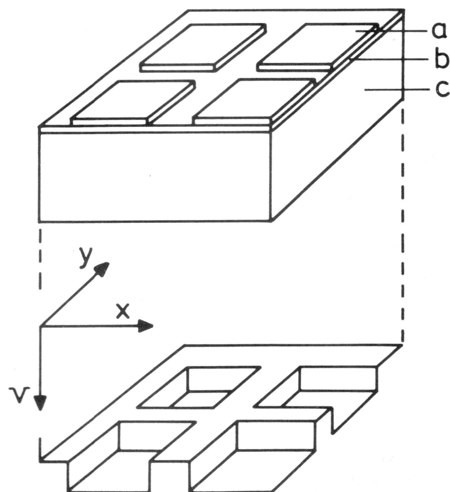


Ir. M.G. Collet (38) is natuurkundig ingenieur (Delft, 1966). Sinds zijn afstuderen werkt hij op het natuurkundig laboratorium van Philips. Tot 1972 bestudeerde hij het siliciumoxyde-siliciumgrensvlak en de eigenschappen van kristalfouten in silicium.
Daarna hield hij zich bezig met de 'device' fysica van ladingsoverdracht-inrichtingen (CTD's). Vanaf 1977 maakt hij deel uit van een werkgroep die het realiseren van een vastestofbeeldopnemer tot taak heeft.

Het vastleggen van bewegende beelden blijft veel mensen boeien. Het ligt voor de hand dat de elektronische industrie zou proberen met elektronische middelen optische beelden weer te geven. Dit heeft geresulteerd in televisie en videorecording. De volgende stap is de ontwikkeling van een handzame goedkope TV-camera. Een essentieel onderdeel hiervoor is een vastestofbeeldopnemer.
Sinds de opkomst van de geïntegreerde circuits is gezocht naar een circuit waarmee een beeld omgezet kan worden in een elektronisch signaal. Dat circuit zou veel goedkoper moeten zijn dan de opneembuizen gebruikt in studiocamera's. Het zou bovendien veel kleiner en betrouwbaarder kunnen zijn. Nieuwe principes (emmertjesgeheugen[1]), ladingsgekoppelde inrichting[2] en de versnelde technologische ontwikkeling ('very large scale integration') hebben de vastestofbeeldopnemer en daarmee een goedkope TV-camera *bijna* mogelijk gemaakt.
In deze bijdrage worden kort enkele ontwerpen van beeldopnemers geschetst. Daarna volgt een iets uitgebreider beschrijving van de beeldopnemer met weerstandselektroden waaraan op ons laboratorium gewerkt wordt.

Integratie van fotostroom, lekstroom
Door het aanleggen van een positieve spanning aan het oppervlak van een p-geleidend siliciumplaatje kunnen de gaten van het oppervlak worden weggeduwd. Zo ontstaat een laagje waaruit alle gaten verwijderd zijn. Deze depletielaag heeft een dikte van enkele microns (10^{-6} m). Hierin heerst een veld dat elektronen de laag intrekt en gaten wegduwt. Wanneer door invallend licht gat-elektronparen gegenereerd worden zal het veld de gaten en elektronen van elkaar scheiden. Zo ontstaat een fotostroom.
In een vastestofbeeldopnemer met de resolutie van een TV-beeld moet de fotostroom op ongeveer 250.000 plaatsen gemeten worden. Door de fijne onderverdeling wordt de fotostroom per beeldpunt zeer klein. Daarom wordt de fotostroom in ieder beeldpunt enige tijd geïntegreerd. Zo ontstaan ladingspakketjes die door het silicium of via geleidersporen op het silicium naar een uitgangsversterker worden getransporteerd. Voor het integreren van de fotostroom zijn – zeer kleine – condensatoren nodig. Een hiervoor veel gebruikte structuur (figuur 1) bestaat uit metaalvlekjes op een isolerende siliciumdioxyde huid op het p-type kristal. Een positieve spanning op de vlekjes zorgt voor lokale depletiegebieden. Hierin worden de gegenereerde elektron-gatparen van elkaar gescheiden. De foto-elektronen worden verzameld op de vlakke-plaatcondensator gevormd door de metaal-oxyde-silicium (MOS)-structuur. Het transport van de ladingspakketjes naar de uitgang vormt een hoofdstuk apart. Gat-elektronparen kunnen behalve door licht ook door thermische excitatie worden gegenereerd. Deze thermische generatie wordt door bepaalde verontreinigingen en kristalfouten enorm versneld. Om te voorkomen dat deze lekstroom het beeld bederft moeten zeer hoge eisen aan de siliciumtechnologie worden gesteld. Op 10^{12} siliciumatomen mag er slechts één ongewenst verontreinigingsatoom aanwezig zijn. De gemiddelde concentratie kristalfouten mag niet veel groter zijn dan één per vierkante centimeter.

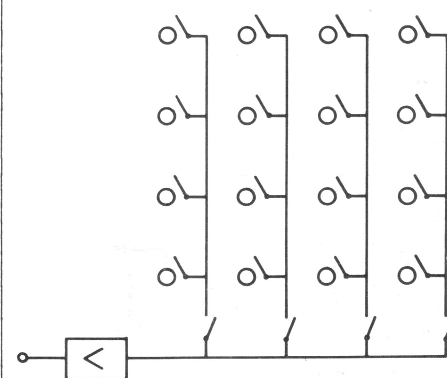
Fig. 1. Matrix van metaal-oxyde-silicium-condensatoren en het potentiaalprofiel dat bij het siliciumoppervlak wordt geïnduceerd. a: metaalvlekje; b: siliciumdioxyde; c: p-type silicium.



Een uiterst summiere schets van drie door veel laboratoria intensief onderzochte beeldopnemers
De *X-Y opnemer* is het oudste type opneemcircuit (figuur 2). Het maakt gebruik van een schakelmatrix. Elk beeldelement wordt op zijn beurt galvanisch verbonden met een verticale Y-lijn en een horizontale X-lijn en via deze weg met een uitgangsversterker. Voor de schakelaars worden MOS-transistoren gebruikt. Het grote voordeel van deze X-Y-opnemer is dat hij in een al op grote schaal toegepaste technologie kan worden gemaakt. Met de X-Y-opnemer worden op het moment dan ook de beste beelden gemaakt[3]. Het grote nadeel is de capaciteit van de verticale en horizontale geleidersporen. Deze is veel groter dan de capaciteit van het beeldelement. Het kleine pakket foto-elektronen geeft maar een heel kleine potentiaalverandering op de geleiders. Tengevolge van deze verzwakking is dit soort opnemer erg ongevoelig.

Dit bezwaar kan worden ondervangen door gebruik te maken van 'charge coupled devices' (CCD). Een CCD (figuur 3a) kan bestaan uit een n-type laag op een p-type kristal. De n-type laag is bedekt met een dunne oxydelaag. Hierop zijn elektroden aangebracht. De n-laag is zo dun en licht gedoteerd dat alle elektronen

Fig. 2. Basisschema van X-Y-beeldopnemer.



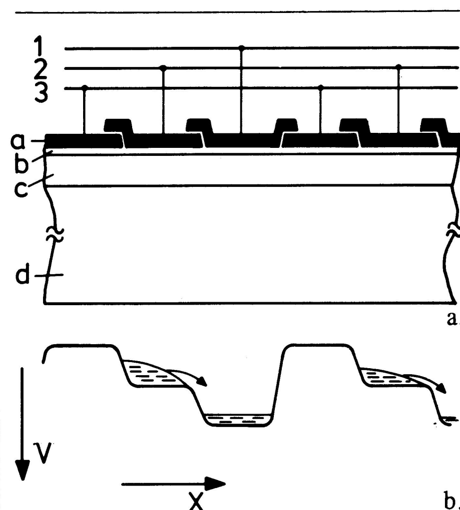
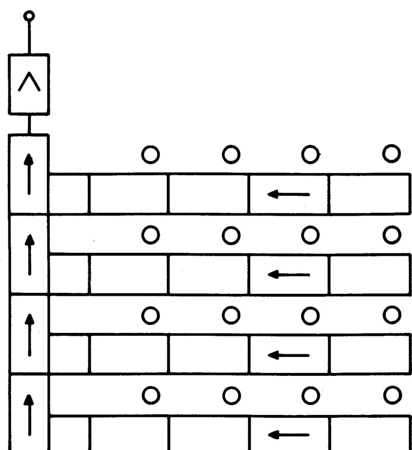


Fig. 3a. Dwarsdoorsnede door CCD. a: elektrode van polykristallijn silicium; b: siliciumdioxyde; c: n-type silicium; d: p-type silicium.
b. Potentiaalverloop en ladingsoverdracht in de n-laag van figuur 3a.

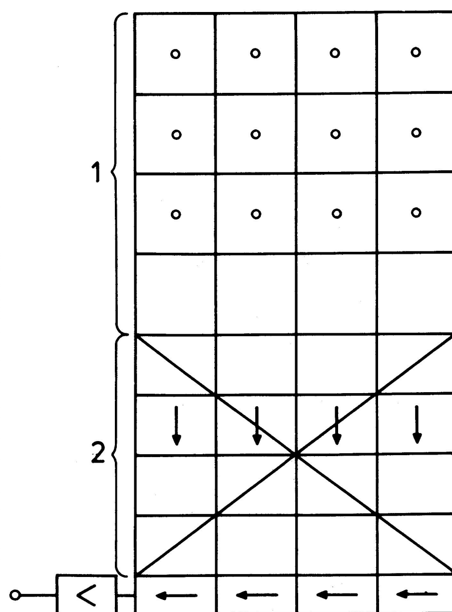
door het aanleggen van een positieve spanning aan deze n-laag kunnen worden verwijderd. Daarna worden de elektroden op verschillende spanningen gebracht. Hierdoor ontstaan in de n-laag potentiaalmaxima, potentiaalputten voor elektronen. In de potentiaalputten kunnen pakketjes elektronen, geïsoleerd van elkaar, worden opgeslagen. Als elke CCD-cel drie elektroden heeft kan één elektrode (op negatieve potentiaal) de pakketjes gescheiden houden. Een positieve elektrode zorgt voor een ontvangende potentiaalput, een van positief naar negatief geschakelde stuwt het pakket voort (figuur 3b). Door dit proces van elektrode tot elektrode te herhalen kunnen pakketjes elektronen door een siliciumlaag worden getransporteerd. De verliezen zijn laag, minder dan een duizendste procent per stap, het transport is bovendien snel.
In de zogenaamde *'interline transfer'*[4] *beeldopnemers* (figuur 4) bevindt zich naast elke kolom beeldelementen een CCD-transportregister. Gedurende de integratietijd wordt in elk beeldelement een pakket foto-elektronen gevormd. Aan het eind van de integratietijd worden alle ladingspakketten in de verticale CCD-registers geplaatst. Deze sluiten aan op een gemeenschappelijk horizontaal CCD. Dit transporteert telkens één lijn informatie naar de uitgangsversterker en ontvangt dan weer een nieuwe lijn uit de verticale registers. In de tijd nodig voor het uitlezen van de registers worden in de beeldpunten weer nieuwe ladingspakketten gevormd. Tijdens het transport mogen de ladingspakketten niet meer door opvallend licht worden vergroot. Daarom worden de verticale registers bedekt met ondoorzichtig materiaal. Dit betekent dat een flink deel van het opvallende licht niet wordt gebruikt. Bovendien wordt de horizontale afstand tussen de beeldelementen vergroot door de ruimte vereist voor de verticale registers. Dit beperkt het scheidend vermogen dat met een circuit van gegeven afmetingen bereikbaar is.

Fig. 4. 'Interline transfer' beeldopnemer.



In de *'frame transfer' opnemer*[5] wordt de fotostroom geïntegreerd in de potentiaalputten van een CCD-register (figuur 5). Het lichtgevoelige deel van dit circuit bestaat uit parallelle CCD-registers. Gedurende de integratietijd worden de elektrodepotentialen constant gehouden. Vervolgens wordt het hele ladingspatroon snel naar een buffergeheugen getransporteerd. Ook dit buffergeheugen bestaat uit parallelle CCD-registers maar afgeschermd tegen licht. Het aantal cellen per register moet hier even groot zijn als in het lichtgevoelige deel. Maar het oppervlak per cel kan kleiner zijn, omdat de opslag van lading minder oppervlak vraagt dan de generatie. De uitlezing van het buffergeheugen is analoog aan de uitlezing van de 'interline transfer' beeldopnemers.
Voor het lichtgevoelige deel van de frame transfer-schakeling zijn transparante elektroden nodig. Polykristallijn silicium voldoet heel goed als elektrodemateriaal en laat groen en rood licht door. Een kleurencamera moet ook gevoelig zijn voor blauw licht. Indium- en tinoxyde lijken hiervoor bruikbaar maar kunnen nog niet in de siliciumtechnologie worden ingepast.

Fig. 5. 'Frame transfer' beeldopnemers.
1: lichtgevoelig gedeelte;
2: buffergeheugen.



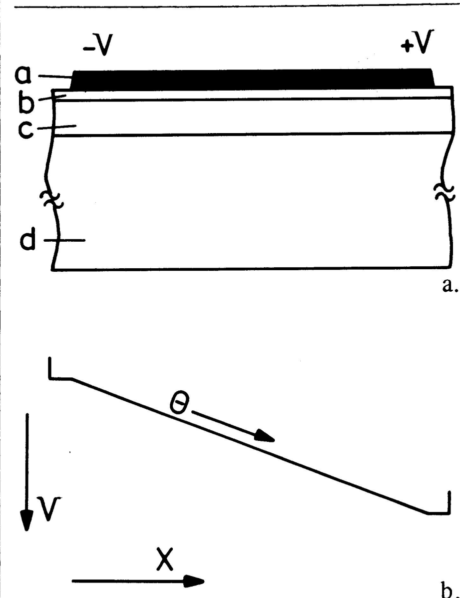


Fig. 6a. Dwarsdoorsnede door transportkanaal met weerstandselektrode. a: hoogohmige elektrode; b: siliciumdioxyde; c: n-type silicium; d: p-type silicium.
b. Potentiaalverloop en ladingsoverdracht in de n-laag van figuur 6a.

Beeldopnemer met weerstandselektroden
We willen een gevoelige beeldopnemer, dus geen X-Y-uitlezing. We willen ook blauw licht detecteren, dus niet het geheel met elektroden bedekte oppervlak van frame transfer. Resteert de interline transfer-organisatie, maar liefst met smalle banen voor het verticale transport.
Een nieuw transportmechanisme lijkt hiervoor mogelijkheden te geven (figuur 6): een CCD-structuur met slechts één elektrode. De elektrode is gemaakt van hoogohmig polykristallijn silicium, zodat een flink spanningsverschil tussen de uiteinden weinig dissipatie geeft. Net als in het CCD is de n-laag en een deel van het onderliggende p-materiaal gedepleerd. Het spanningsverschil langs de elektrode induceert een gradiënt in de n-laag (figuur 6b). Rekening houdend met de vertragende effecten van diffusie en coulombafstoting kan men berekenen dat een elektronenpakket in $25 \mu s$ langs een 0,4 cm lang kanaal kan worden getransporteerd met een potentiaalverschil van 10 V. Het tijdschema gebruikt bij televisie- en video-

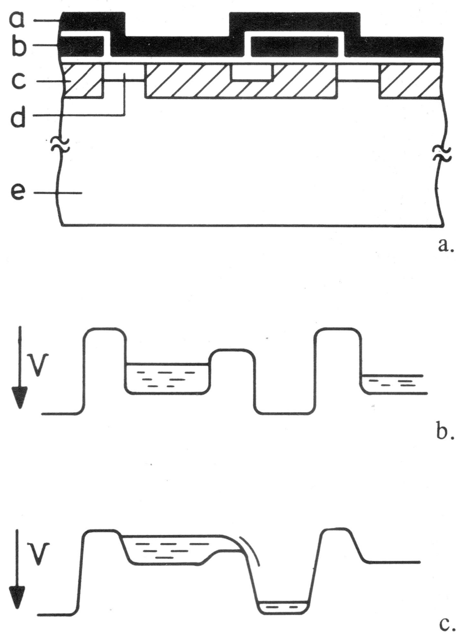
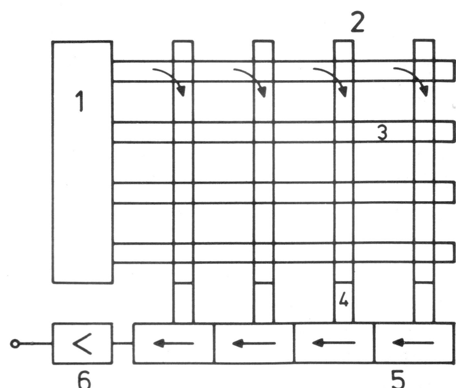


Fig. 7a. Dwarsdoorsnede door beeldelement en transportkanaal van een beeldopnemer met weerstandselektroden.
a: integratie-elektrode; b: weerstandselektrode; c: n-type silicium; d: zwaar gedoteerd p-type silicium; e: p-type silicium.
b. Potentiaalverdeling tijdens integratie.
c. Potentiaalverdeling bij ladingsoverdracht.

recording staat 64 μs toe. Dit mechanisme is dus voldoende snel voor het verticale transport van de ladingspakketten, mits we voldoende beeldlijnen op 0,4 cm kunnen plaatsen.
De koppeling van de beeldelementen met het transportkanaal is geschetst in figuur 7. In één n-gebied worden beeldelement en transportkanaal gemaakt. Ze zijn van elkaar gescheiden door een ondiep zwaar gedoteerd p-gebied. De potentiaal in het gedepleerde n-gebied onder de p-baan wordt door de vaste potentiaal van de p-baan op een kleine positieve waarde gehouden. De potentiaal van het integratiegebied wordt bepaald door de integratie-elektrode. Tijdens de integratie is deze flink positief zodat een elektronenpakket kan worden verzameld. Daarna wordt de elektrode naar nul geschakeld. De potentiaalput wordt ondieper, de elektronen worden geïnjecteerd in het transportkanaal (figuur 7c). De andere begrenzing van het integratiegebied wordt gevormd door een zwaar gedoteerd p-gebied in direct contact met het p-kristal. De potentiaalbarrière aan die kant is zo hoog dat de elektronen alleen in het 'eigen' transportkanaal geïnjecteerd worden.
Met deze elementen hebben we de beeldopnemer met weerstandselektroden ontworpen (figuur 8)[6]. De transportkanalen lopen verticaal, de integratie-elektroden horizontaal. Met een digitaal schuifregister kunnen de integratie-elektroden op eenvoudige wijze sequentieel naar nul volt worden geschakeld. De geïnjecteerde ladingspakketten worden na transport onder een bufferelektrode verzameld. Tijdens het transport wordt een voorgaande lijn uitgelezen door het horizontale CCD-register. Als dit register leeg is geeft de bufferelektrode een nieuwe lijn over aan het CCD. Zo kan de hele lijntijd worden gebruikt voor het verticale transport.
De elektroden laten een flink deel (25%) van het oppervlak onbedekt. Dit type sensor heeft dus een behoorlijke gevoeligheid voor blauw licht.

Fig. 8. Beeldopnemer met weerstandselektroden. 1: digitaal register; 2: verticale transportbaan; 3: integratie-elektrode; 4: bufferelektrode; 5: CCD-register; 6: uitgangsversterker.



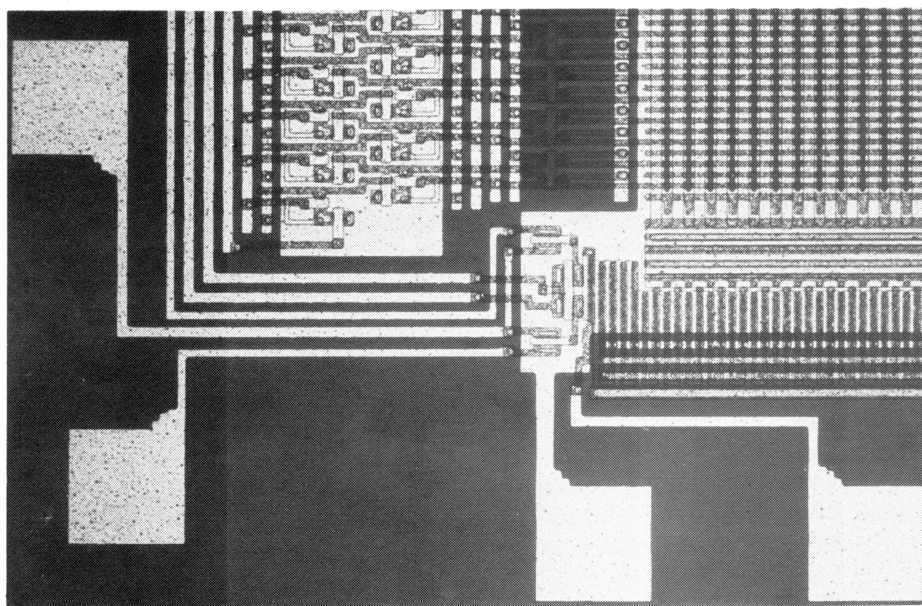


Fig. 9. Microscoopopname van een gedeelte van de beeldopnemer met weerstandselektroden.

Bereikte resultaten
Om onze ideeën te testen hebben we een opnemer ontworpen met 300 lijnen en 200 kolommen. Het beeldformaat is 0,42 cm bij 0,56 cm. Het totale circuit meet 0,55 cm bij 0,75 cm.
Een microscoopopname (figuur 9) toont rechts de verticale transportkanalen met

Fig. 10. Beeld opgenomen met het circuit van figuur 9.



loodrecht daarop de integratie-elektroden, links daarvan het digitale schuifregister voor de lijn-aansturing. Onder de transportkanalen het CCD-register, links aansluitend op de uitgangsversterker. De bereikte beeldkwaliteit is behoorlijk (figuur 10). Met 50 000 elektronen per signaalpakket wordt een signaal-ruisverhouding van 40 dB bereikt. Om elektronenpakketjes van deze grootte te genereren is, bij gebruik van een F2-lens, een verlichtingssterkte van 120 lux voldoende. Dit komt overeen met het lichtniveau in een huiskamer.
Voor voldoende horizontale resolutie is een groter aantal beeldpunten per lijn nodig. Wanneer we bovendien kleur- en helderheidsinformatie met één opnemer willen verkrijgen, dan moeten we naar 600 lijnen van elk 400 kolommen. Dat vraagt grotere circuitafmetingen en dus moeten de verticale transportkanalen met hun weerstandselektrode ook langer worden. Om na te gaan hoever we kunnen gaan hebben we de transportsnelheid van de huidige kanalen gemeten. Daartoe injecteren we de bovenste lijn en meten we de stroom aan het eind van de 0,42 cm lange kanalen als functie van de tijd (figuur 11). De getrokken lijn geeft het theoretisch verwachte resultaat voor een aangelegd potentiaalverschil van 10 V en een elektronenbeweeglijkheid van 750 $cm^2 \cdot V^{-1} \cdot s^{-1}$. De onderbroken lijnen geven de meetresultaten voor toenemende grootte van de ladingspakketten. Voor grote pakketten komt ongeveer de helft van de lading met de verwachte tijdsverdeling naar buiten. De rest van de lading is verdeeld over een relatief lange staart. De staart verzadigt met toenemende pakketgrootte. Het lijkt erop dat een deel van de elektronen wordt ingevangen en met variërende tijdconstante weer wordt vrijgelaten. Het aantal invangplaatsen is beperkt, gezien de verzadiging van de staart met toenemende pakketgrootte.

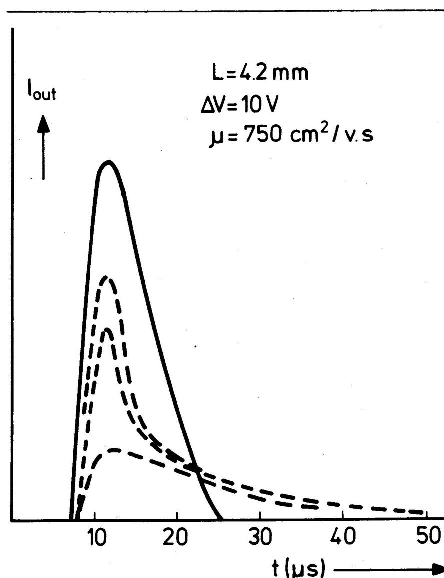


Fig. 11. De uitgangsstroom (in arbitraire eenheden) van de verticale transportkanalen als functie van de tijd na injectie van de bovenste rij ladingspakketten.

Het is bekend dat in silicium verontreinigingen aanleiding kunnen geven tot gelokaliseerde energieniveaus tussen geleidings- en valentieband. Dergelijke niveaus kunnen elektronen invangen en – door thermische excitatie – weer vrijgeven. Om de waargenomen effecten te verklaren zijn ongeveer 5 x 10^{12} van deze centra per kubieke centimeter silicium nodig. Het is ook mogelijk dat variaties in doteringsconcentratie of afmetingen van het kanaal kuilen veroorzaken in het potentiaalverloop langs het kanaal. Als daardoor gelokaliseerde potentiaalputjes ontstaan, dan kunnen daarin elektronen worden ingevangen en door thermische excitatie weer worden vrijgemaakt. Nader onderzoek zal ons moeten leren welk van deze effecten ons parten speelt en welke tegenmaatregelen kunnen helpen. Pas daarna zullen we de beeldopnemer significant kunnen vergroten.

Slotopmerkingen
Dit verhaal laat iets zien van de mogelijkheden voor het ontwerp van vastestof-beeldopnemers. Daarbij ontbrak de ruimte om op deelproblemen in te gaan: hoe verdelen de foto-elektronen zich over de verschillende beeldelementen, hoe voorkom je het overlopen van overbelichte elementen, welke kleurfilterontwerpen zijn optimaal, welke bewerkingen moet je nog op het uitgangssignaal toepassen, hoe kun je de optimale doteringsconcentraties en laagdiktes berekenen. De opgegeven litteratuurplaatsen geven daar gedetailleerde informatie over.
We moeten hier volstaan met de conclusie dat in verschillende laboratoria is aangetoond dat met vastestofbeeldopnemers goede kleurencamera's gemaakt kunnen worden. Na verdere perfectionering van de siliciumtechnologie zal het ook mogelijk worden de opnemers foutvrij en goedkoop te maken.

Referenties

1. F.L.J. Sangster, K. Teer: *Bucket-Brigade Electronics*. IEEE Journal on Solid State Circuits **SC-4** (1969) 131.
2. W.S. Boyle, G.E. Smith: *Charge Coupled Semiconductor Devices*. Bell Syst. Techn. J. **49** (1970) 587.
3. Norio Koike, Iwao Takemoto, Kazuhiro Sato, Haruo Matsumaru, Mikio Ashikawa, Masaharu Kubo: *An npn Structure 484 x 384 MOS Imager for a Single Chip Color Camera*. IEEE International Solid-State Circuits Conference 1979, 192.
4. en 5. David F. Barbe: *Imaging Devices Using the Charge-Coupled Concept*. Proc. of the IEEE **63** (1975) 39.
6. H. Heyns, H.L. Peek, J.G. van Santen: *Beeldsensor met weerstandselektroden*. Philips Techn. T. **37** (1977) 321.

De ontwikkeling van ionengevoelige veldeffecttransistoren (ISFET) op basis van de MOS-IC-technologie

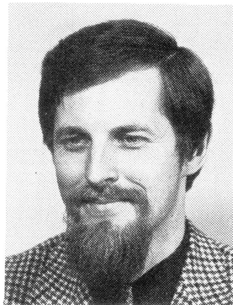
P. Bergveld en N.F. de Rooij

De ISFET (ion-sensitive field effect transistor) is een ionengevoelige sensor, ontwikkeld door integratie van de ionengevoelige glaselektrode en de MOSFET (metal-oxide-silicon field effect transistor). Werking, vervaardiging en toepassing van de ISFET worden beschreven.

Historie
Reeds uit de jaren twintig stamt het idee de elektrische geleiding van een plaatje halfgeleidermateriaal te veranderen met een elektrisch veld, loodrecht op het plaatje. Het zou daardoor mogelijk zijn de elektrische stroom *energieloos* te sturen in een kleine robuuste structuur, zonder gebruikmaking van een gloeidraad.[1]

Uitvoerige metingen van Bardeen en Brattain van Bell Telephone Labs. in de periode 1947-50 aan germanium-éénkristallen leidden niet tot het genoemde veldeffect. De oorzaak lag in het feit, dat de geïnduceerde ladingsdragers immobiel werden door vangst in vrije valentiebindingen: de *oppervlaktetoestanden*. Het bewijs, dat de oppervlaktetoestanden kunnen worden geneutraliseerd, leverden Garrett en Brattain bij metingen aan een germanium/elektrolytcel. Ook bleek dat onder kathodische voorspanning de geleiding van het germanium aanzienlijk kon worden beïnvloed. Praktische betekenis hadden deze experimenten niet. Wel brachten deze onderzoeken Shockley tot de uitvinding van de *pn-overgang* (1949) en de *junctietransistor* (1950).

In de jaren die volgden onderging de junctietransistor een grote ontwikkeling. De oppervlaktetoestanden bleven echter van storende invloed op de eigenschappen van de transistor. Veel werk werd gestoken in



Dr. ir. P. Bergveld (39) studeerde elektrotechniek aan de TH-Eindhoven. In 1965 trad hij in dienst bij de Technische Hogeschool Twente als medewerker van de vakgroep Bio-informatica. Hij heeft zich daar gespecialiseerd in het ontwikkelen van meet- en stimulatiemethoden ten behoeve van de biomedische research. Zijn speciale aandacht heeft de koppeling tussen de 'ionenwereld' en de 'elektronenwereld' onder andere via de in dit artikel beschreven ionengevoelige veldeffect-transistor. Op dit onderwerp is hij in 1973 gepromoveerd bij prof. ir. M.P. Breedveld en prof. ir. O.W. Memelink. Hij geeft college in de biomedische instrumentatie en is projectleider van de werkgroep Biosensoren.

de beheersing van deze toestanden. Inmiddels begon silicium germanium te verdringen als uitgangsmateriaal voor de transistorfabricage. Een beslissende stap werd gedaan door Atalla (1958), die ontdekte dat een thermisch gegroeide SiO_2-laag in staat is het grootste deel van de oppervlaktetoestanden te neutraliseren. Snel werden toen ook de basisprincipes van de *planaire siliciumtechniek* vastgelegd, gevolgd door de uitvinding van de MOSFET door Kahng en Atalla (1960).

Werking van MOSFET en CHEMFET
De MOSFET heeft de volgende opbouw (figuur 1). Een p-type siliciumplaatje (*chip*) bevat twee n-gebiedjes (*source en drain*). Tussen source en drain bevindt zich een MOS-capaciteit met een dikte van het diëlektricum (SiO_2) van ca. 0,1 μm. Een positieve spanning op het metaal (*gate*) ten opzichte van het silicium induceert inversie van het geleidingstype van het silicium waardoor een elektronenkanaal ontstaat dat een geleidende verbinding tussen source en drain vormt. De kanaalgeleiding wordt gegeven door:

$$g = \mu C_{ox}(W/L)(V_g - V_t) \qquad (1)$$

waarin μ de beweeglijkheid der elektronen is, C_{ox} de MOS-capaciteit per oppervlakte-eenheid, L de lengte van het kanaal tussen source en drain, W de breedte van het kanaal, V_g de gate-potentiaal en V_t de *drempelspanning* is. Uitdrukking (1) geldt alleen wanneer V_g de waarde V_t overschreden heeft en inversie van het geleidingstype optreedt.

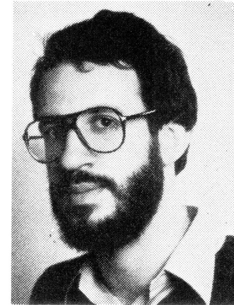
De drempelspanning V_t blijkt chemisch beïnvloedbaar te zijn. V_t kan worden geschreven als:

$$V_t = \Phi_m - \Phi_s - (Q_{ss} + Q_{ox})/C_{ox} + + 2\Phi_f - Q_B/C_{ox} \qquad (2)$$

Φ_m en Φ_s zijn resp. de uittreepotentialen van gate-metaal en silicium. $(Q_{ss} + Q_{ox})/C_{ox}$ is de bijdrage tot V_t veroorzaakt door lading Q_{ss} in oppervlaktetoestanden en lading Q_{ox} in het oxyde, die eveneens aan het grensvlak Si/SiO_2 gedacht is. De eerste drie termen van (2) geven de *vlakkebandenspanning* V_f:

$$V_f = \Phi_m - \Phi_s - (Q_{ss} + Q_{ox})/C_{ox} \qquad (3)$$

V_f is de gatespanning waarbij de elektronenenergiebanden in de halfgeleider tot aan het oppervlak *vlak* zijn.[2] De laatste twee termen van uitdrukking (2) stellen de extra gate-spanning voor die nodig is om het halfgeleideroppervlak te inverteren. Φ_f is de zg. fermipotentiaal bepaald door de kristaldotering en Q_B is de lading per oppervlakte-eenheid in de verarmingslaag onder het oppervlak, die eveneens van de dotering afhankelijk is.
Chemische sensoren (*CHEMFET's*) berustend op het MOS-systeem, opereren met het oog op het te gebruiken meetprincipe met een zodanige lading Q_{ox} dat een inversielading $-Q_{ox}$ wordt geïnduceerd, zodat voor $V_g = 0$ reeds kanaalgeleiding aanwezig is. Door chemische beïnvloeding wordt, onder de conditie $V_g = 0$, Φ_m gevarieerd, wat zich uit in een verandering van de kanaalgeleiding. Voor een goed overzicht van soorten CHEMFET's wordt verwezen naar Croset.[3] In het volgende zullen wij ons in het bijzonder bezighouden met de *ISFET (ion-sensitive field effect transistor)*. Voor een overzicht van recent ISFET-werk wordt verwezen naar Janata en Huber.[4]



Dr. N.F. de Rooij (28) studeerde scheikunde aan de RU-Utrecht. Na zijn afstuderen trad hij in ZWO-dienst om aan de TH-Twente een promotieonderzoek te verrichten naar de fysische en chemische verschijnselen, die ten grondslag liggen aan de werking van de ionengevoelige veldeffecttransistor (ISFET). In 1978 werd dit onderzoek afgesloten met een promotie. De promotoren waren prof. dr. J.Th.G. Overbeek en prof. dr. M.J. Sparnaay.

Theorie van de ISFET
Om tot een ionengevoelige structuur te komen wordt het gate-metaal van de MOSFET van figuur 1 vervangen door een elektrolytoplossing nadat source- en draincontacten genoegzaam zijn afgedekt. Via een referentie-elektrode kunnen wij de elektrolytoplossing een bepaalde spanning geven ten opzichte van het silicium (figuur 2).
In de ESO (elektrolyt-oxyde-silicium)-structuur wordt de vlakkebandenspanning gegeven door

$$V_f = E_r - {}^{el}\Delta^{ox}\Phi - \Phi_s - (Q_{ox} + Q_{ss})/C_{ox} \qquad (4)$$

waarin E_r de referentie-elektrodepotentiaal voorstelt en ${}^{el}\Delta^{ox}\Phi$ het elektrisch potentiaalverschil tussen het oxydeoppervlak en de elektrolytoplossing. Tengevolge van de interactie met water ont-

Fig. 1. Doorsnede van de MOSFET.

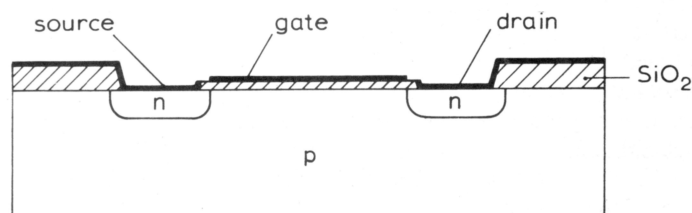
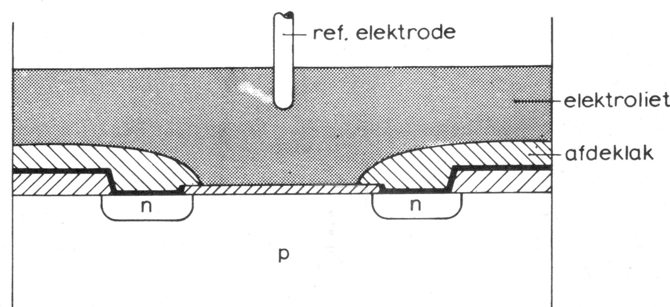


Fig. 2. ISFET met elektrolyt en referentie-elektrode.



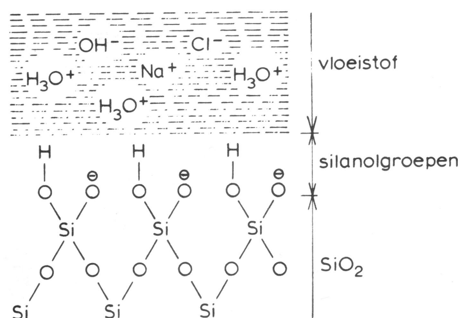
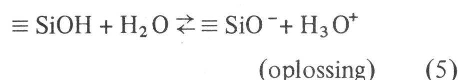


Fig. 3. Gehydrateerd oxyde-oppervlak met silanolgroepen.

staan silanolgroepen aan het oppervlak die gedeeltelijk gedissocieerd zijn volgens het evenwicht



$$\equiv \text{SiOH} + H_2O \rightleftarrows \equiv \text{SiO}^- + H_3O^+ \quad \text{(oplossing)} \qquad (5)$$

Schematisch is dit evenwicht weergegeven in figuur 3. Dit betekent dat het SiO_2-oppervlak elektrisch geladen zal zijn, resulterend in een potentiaalverschil ${}^{el}\Delta^{ox}\Phi$ ten opzichte van de oplossing. Daarom zal de H_3O^+-concentratie in de oplossing, of m.a.w. de *zuurgraad (pH)*, van invloed zijn op de oppervlaktelading en dus ook op ${}^{el}\Delta^{ox}\Phi$. In de praktijk blijkt dat ${}^{el}\Delta^{ox}\Phi$ van meer ionsoorten dan alleen H^+-ionen afhankelijk is, zoals bijvoorbeeld van Na^+- en K^+-ionen.

Vervaardiging van de ISFET

De voor de vervaardiging toegepaste *planair silicium techniek* heeft zijn voordelen reeds bewezen bij de realisatie van *geïntegreerde schakelingen (IC's)*. Typerend voor deze techniek is de massafabricage van zeer kleine gelijkvormige elektronische componenten of systemen. Dit betekent dat de ISFET eventueel met de elektronische signaalverwerking op één chip kan worden geïntegreerd.

De op de TH-Twente vervaardigde ISFET's zien eruit als gegeven in figuur 4. Tijdens het thermisch groeien van het gateoxyde ontstaat door positieve elektronenvacatures in het overgangsrooster $Si/SiO_x/SiO_2$ een *positieve* oxydelading $Q_{ox} + Q_{ss}$. Met het oog op de reeds genoemde eis, dat inversie aanwezig moet zijn voor $V_g = 0$ is gekozen voor een *n-kanaal*-ISFET.

De vervaardiging van de ISFET geschiedt als volgt. Op een siliciumplak van 3 cm doorsnee waarin vele gelijkvormige ISFET-structuren zullen ontstaan wordt door middel van *chemical vapour deposition (CVD)* een met borium gedoteerde SiO_2-laag (SiO_2(B)) van 0,5 μm dikte gedeponeerd. Fotolithografisch wordt dit oxyde verwijderd daar waar later source en drain van de ISFET komen. Figuur 4a geeft een bovenaanzicht waarin de SiO_2(B)-laag is gearceerd terwijl figuur 4b een doorsnede door de chip laat zien langs de lijn A-B.

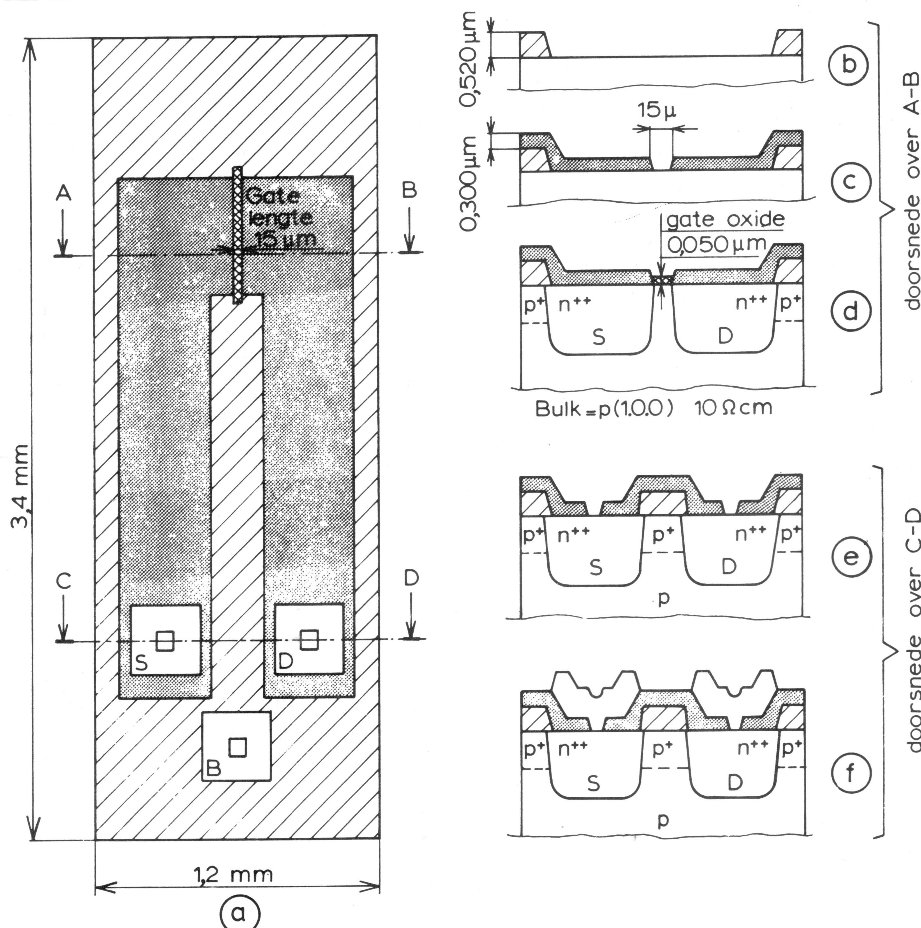


Fig. 4. a: bovenaanzicht ISFET-chip. b, c, d: doorsneden over AB. e, f: doorsneden over CD.

Hierna wordt fosforgedoteerd oxyde door middel van CVD over de hele plak gedeponeerd. Een tweede fotolithografie stap opent het oxyde daar waar het ionengevoelige oxyde komt (figuur 4c). Dit geschiedt door verhitten van de plak in een zuurstofatmosfeer en daarna stikstofbij een temperatuur van 1200 °C. Simultaan treden de volgende effecten op (figuur 4d):

– groei van 0,05 μm gateoxyde;
– vorming van source en drain met lange n-type geleidende banen (eigen weerstand ca. 100 Ω) door fosfordiffusie vanuit het oxyde;
– verhoging van acceptordotering buiten de eigenlijke ISFET-structuur door boriumdiffusie vanuit het oxyde waardoor parasitaire inversielagen worden onderdrukt.

Vervolgens worden de contactgaten in het fosforgedoteerde oxyde geëtst (figuur 4e). Opdampen van aluminium en fotolithografisch etsen verschaft de contactvlakjes voor source, drain en bulk. De plak met de complete ISFET's wordt dan gekrast en gebroken in individuele chips. Iedere chip wordt op een stripvormige Al_2O_3-drager bevestigd. Bovendien worden de drie contactvlakjes met dunne gouddraadjes met de geleidingssporen op het Al_2O_3-substraat verbonden.

Tot besluit wordt de ISFET geheel met siliconenrubber afgedekt, behalve het gategebied (figuur 5).

Meetresultaten

In principe berust de ionenconcentratiemeting met de ISFET op het meten van de kanaalgeleiding bij een gegeven referentie-elektrodepotentiaal ten opzichte van de source. De gemeten ionenconcentraties kunnen worden weergegeven als een equivalente gatespanning $V_{g,eq}$. Een voorbeeld geeft figuur 6 waaruit blijkt, dat een thermisch oxyde zonder verdere bedekking gevoelig is voor H^+-, Na^+- en K^+-ionen.

Om de selectiviteit van een ISFET te bevorderen zijn experimenten gedaan met een laagje parafine, gemengd met koolstof en chloorgevoelig Ruzicka-poeder, aangebracht op het gateoxyde. Inderdaad bleek de verkregen sensor selectief te zijn voor halogenide-ionen met een perfecte *Nernst-helling* van 59 mV per decade voor Cl^--ionen (figuur 7).

Veel onderzoek is momenteel gaande naar laagjes of membranen met bijzondere selectieve eigenschappen. Behalve dat op de beschreven wijze bestendige en kleine ionengevoelige sensoren zijn te vervaardigen moet nog worden vermeld dat ook veel aandacht is besteed aan elektronische

meetmethoden waarbij storende invloeden als temperatuurgevoeligheid worden gecompenseerd.[5]
Dergelijke ISFET-systemen zijn toegepast voor snelle zuur-basetitraties. Het blijkt dat bij titratiesnelheden van 5 ml/minuut, het ISFET-meetsysteem afwijkingen kleiner dan 1% oplevert, terwijl de normale glaselektrode door zijn traagheid slechts bij titratiesnelheden kleiner dan 0,9 ml/minuut binnen de 1% fout blijft. Bij het continu meten van min of meer constante pH-waarden vertoont de ISFET enige meetdrift. Een drift van 1 mV/uur ($\approx$ 0,2 pH-eenheid/uur) is geconstateerd, hoewel in de literatuur ISFET's worden vermeld met een drift van slechts 2 mV per 4 dagen. Diverse technologische procesvariaties worden thans bekeken op hun merites om de drift te verkleinen.

Nabeschouwing
Met de in dit artikel beschreven ontwikkelingen op het gebied van ISFET's is een nieuwe generatie sensoren geïntroduceerd, die evenwel niet alleen staat. Ook ten aanzien van het meten van niet-chemische grootheden (druk, flow, snelheid, versnelling, enzovoort) is een ontwikkeling gaande van microminiaturisatie der transducers, gebruikmakende van IC-technologie.
De ontwikkeling van ISFET's is aanvankelijk gestart ten behoeve van biomedische toepassingen. Er bestaat een grote behoefte aan intracorporale detectie van chemische grootheden, omdat het metabole evenwicht van de patiënt van belang is. Het zal duidelijk zijn dat hiertoe multisensoren dienen te worden ontwikkeld, waarbij de IC-technologie, mogelijk gecombineerd met dunne- en dikke-film-

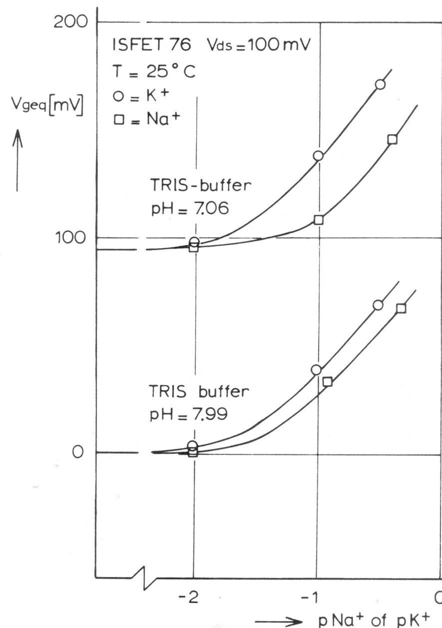


Fig. 6. H^+- K^+- en Na^+-gevoeligheid van ISFET met onbedekt oxyde.

technologie, de aangewezen weg tot realisatie lijkt. De beschreven ISFET is daar nog slechts een eenvoudig voorbeeld van, waarbij echter wel blijkt dat naast de voordelen van de microminiaturisatie ook geheel nieuwe meetconcepten mogelijk worden. De ontwikkeling zal dan ook verder gaan in de richting van transducersystemen in plaats van transducers sec, met automatische parameterbewaking en calibratie.[5] Het ligt voor de hand dat in de toekomst delen van de meet- en proceselektronica ook bij de transducerchip worden ingebouwd. Het lijkt niet uitgesloten dat mettertijd gekozen zal worden voor 'wegwerp' sensoren, een tendens die in de medische wereld hoe langer hoe meer gebruikelijk wordt. Voor een eenmalig gebruik van de sensoren is

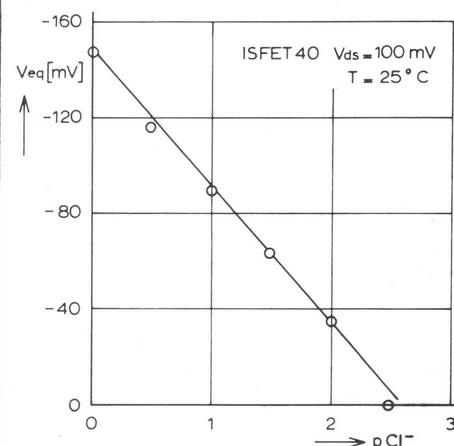


Fig. 7. Cl^--gevoeligheid van halogenide-gevoelige ISFET.

uiteraard massaproduktie noodzakelijk, hetgeen juist door toepassing van de IC-technologie mogelijk wordt, tegen relatief lage kostprijs. De laatste tijd valt ook de belangstelling van de auto-industrie op ten aanzien van het ontwikkelen van vastestofsensoren die bij de carburateur kunnen worden ingebouwd ten behoeve van de ontwikkeling van zuiniger en milieuvriendelijker verbrandingsmotoren. Er zal echter nog een groot aantal problemen dienen te worden opgelost voordat praktisch bruikbare transducersystemen gemeengoed zijn. Afhankelijk van de gebruiksomstandigheden zijn dit problemen op het gebied van de inkapseling der sensoren, de in situ calibratie en een ingebouwd systeem voor een vast referentiepunt dat voor elke meting noodzakelijk is.

De ISFET zoals beschreven in dit artikel is ontwikkeld aan de TH-Twente als vrucht van samenwerking tussen de vakgroepen Bio-informatica, Vastestofelektronica, Vastestoffysica en Chemische Analyse. Behalve de auteurs droegen aan dit artikel bij J. Holleman wat betreft de ISFET-technologie en O.W. Memelink wat betreft de historie van MOSFET-componenten en de eindredactie. Een woord van waardering is op zijn plaats aan A.J. Verloop, die de afmontage van ISFET-chips verzorgt.

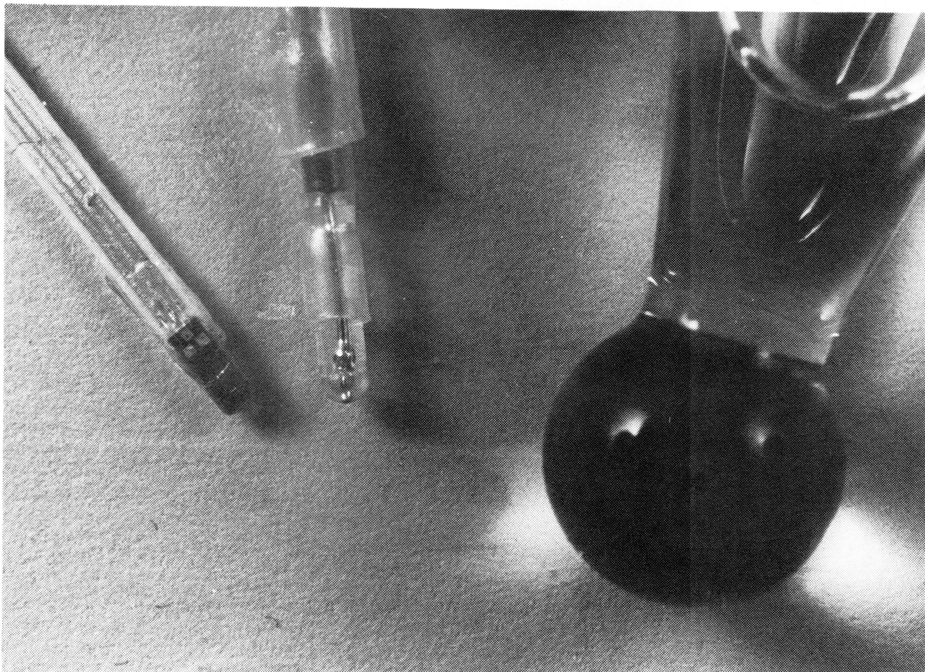


Fig. 5. Van rechts naar links: een conventionele glaselektrodetip, een microglaselektrode, en een ISFET-elektrode.

Referenties
1. Philips Technisch Tijdschrift **31** no 7, 8, 9 (1970), een speciale uitgave over MOS-transistoren.
2. A.S. Grove: *Physics and Technology of Semiconductor Devices*. John Wiley and Sons, New York 1967, p. 278.
3. M. Croset: *Solid State Devices 1978*. ESSDERC 1978, Editions de Physique, Orsay France, p. 47.
4. J. Janata, R.J. Huber, Ion-Selective Electrode Rev. **1** (1979) 31.
5. P. Bergveld, Med. and Biol. Eng. and Computing, 17 (1979) 655. (1979).

De actieve radiozon

A.D. Fokker en J.M.E. Kuijpers.

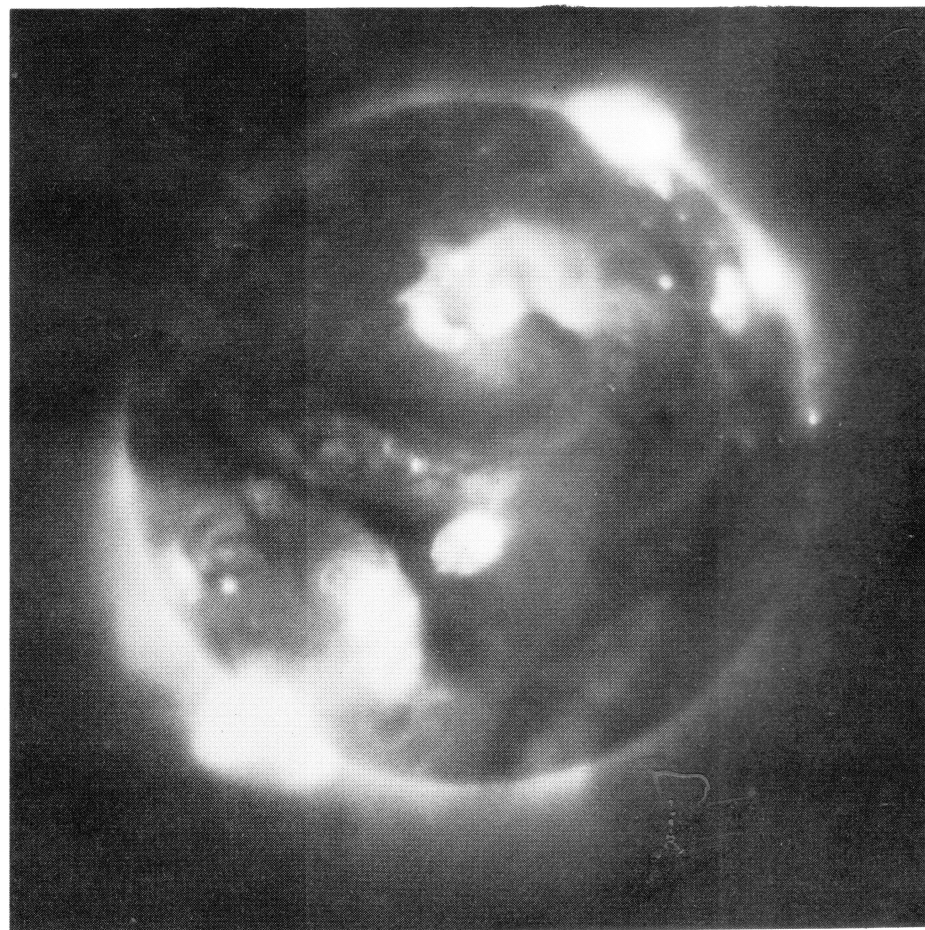
De corona van de zon is een magnetoplasma waarin (bundels van) snelle elektronen instabiele situaties doen ontstaan die leiden tot de opwekking van plasmagolven. Deze geven, volgens verschillende mechanismen, aanleiding tot de opwekking van radiostraling die de corona verlaat. De diverse soorten radio-uitbarstingen of 'transients' zijn gevoelige indicatoren voor de processen die zich in de corona afspelen. Uit kenmerken van die radioverschijnselen kan plausibel worden gemaakt wat voor processen dat zijn.

De titel suggereert al dadelijk dat er verschillende 'zonnen' bestaan. Kenden wij aanvankelijk alleen de optische zon, met zijn scherpe, exact cirkelvormige rand die overeenkomt met het *fotosfeer*-niveau, in de jaren -'40 leerde men de radiozon kennen: veel uitgebreider dan de optische zon en met meer vage en niet cirkelsymmetrische contouren; vervolgens in de jaren -'60 de sterk inhomogene röntgen- (figuur 1) en ver-ultraviolet zon.

De radiostraling van de zon is overwegend afkomstig uit de *corona*, dat is het ijle maar hete en sterk geïoniseerde plasma rondom de zonneschijf waarvan het bestaan vanouds uit totale zonsverduisteringen bekend was.

Opmerkelijk is de abrupte overgang, over een interval van slechts enkele duizenden kilometers, van de relatief koele (T = 5800 K) en dichte ($N \simeq 10^{23}$ m^{-3}) fotosfeer naar de hete (1 à 2 x 10^6 K) en ijle ($N \leqslant 10^{16}$ m^{-3}) corona. De hoge temperatuur van de corona en de daartoe benodigde verhitting wordt in verband gebracht met de aanwezigheid van een circa 200 000 km dikke zone onder de fotosfeer waarin het energietransport hoofdzakelijk door *convectie* plaatsvindt. De zogenaamde *granulatie*, waarmee het zonoppervlak in wit licht overdekt is, is van die convectie een manifestatie en men moet aannemen dat mechanische energie in de vorm van (akoestische en magnetische) golven en schokken aan de corona wordt toegevoerd. Wellicht zijn in het verhittingsmechanisme ook magnetische velden in het spel zoals die veelvuldig door het zonoppervlak uitbreken (waar dat het geval is ontstaat een *zonnevlek* of een kleine zgn. *'pore'*). De mechanische en de eventuele magnetische energieflux vertegenwoordigt maar een kleine fractie (10^{-3}) van de stralingsflux van de zon waarin wij ons plegen te koesteren. Deze magnetische velden vinden hun oorsprong beneden de fotosfeer waar ze door differentiële stromingen worden versterkt. De gasdruk is hier namelijk zo hoog dat de velden praktisch geheel door de thermo- en hydrodynamica van het gas worden bepaald. Omdat een veldbuis steeds in hydrostatisch evenwicht blijft met zijn omgeving wordt de gasdichtheid in de buis na versterking van het veld geringer dan in zijn omgeving. De buis met het lichtere gas rijst nu omhoog, het coronale plasma in. Is in de fotosfeer en daar beneden de gasdruk van overheersend belang, in de corona bepalen daarentegen de magnetische velden de dynamica van het gas. De oprijzende magnetische structuren bereiken hier een nieuw evenwicht. Een deel van de magnetische energie komt daarbij vrij in de vorm van verhitting en versnelling van deeltjes. Op spectaculaire wijze geschiedt dit in een *zonnevlam* (figuur 2).

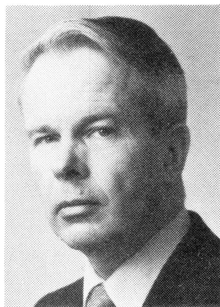
Fig. 1. De inhomogene structuur van de corona is duidelijk te zien op deze röntgenopname in de golflengtebanden 3-32 Å en 44-54 Å, gemaakt op 18 juni 1973 aan boord van Skylab (American Science and Engineering, Cambridge, Mass.).



De zonnevlam

Het meest centrale fenomeen van de gebeurtenissen die zich in een activiteitscentrum afspelen is de zonnevlam met een verscheidenheid van daarmee verbonden effecten (figuur 2), in het bijzonder de radio-uitbarstingen (figuur 3). De zonnevlam ('flare') is een explosief proces waarbij binnen enkele minuten energie beschikbaar komt voor verhitting en deeltjesversnelling (10^{21} – 10^{25} J). Zonder twijfel wordt de energie ontleend aan magnetische energie die tijdens de opbouwfase is opgeslagen. De sterkste vlammen treden bij voorkeur op in activiteitscentra waarin de magnetische configuratie (verdeling van N- en Z-magnetisme in de vlekkengroep en loop van de zgn. 'neutrale' lijn) gecompliceerd is.

Uit het gebied van de primaire energieomzetting ontvangt men op aarde voornamelijk röntgenstraling (remstraling) en radiostraling in het microgolfgebied (gyrosynchrotron- en plasmastraling). De verhitting en versnelling van deeltjes in een zonnevlam leidt verder tot verhitting en 'verdamping' van de dichte fotosfeer aan de voetpunten van de magnetische veldlijnen. De waarnemingen tonen daarbij lussen die afhankelijk van dichtheid



Dr. A.D. Fokker jr. (53) studeerde sterrenkunde te Leiden. Van 1952 tot 1962 was hij werkzaam bij het Staatsbedrijf der PTT in een onderzoeksgroep die de invloed van de zonsactiviteit op de propagatie-condities in de ionosfeer bestudeerde. In 1960 promoveerde hij op een studie naar de eigenschappen van verhoogde radiostraling van de zon. Sinds 1962 is hij verbonden aan de Sterrewacht Sonnenborgh te Utrecht, vanaf 1966 als lector. Zijn onderzoek is gericht op de zonsactiviteit, in het bijzonder de zonne-radioverschijnselen.



Dr. J. Kuijpers (33) studeerde theoretische sterrenkunde aan de RUU. In 1971 trad hij in dienst bij het Sterrekundig Instituut te Utrecht voor theoretisch onderzoek op het gebied van de radiostraling van de zon. In 1975 promoveerde hij op de verklaring van structuurlijke continuümuitbarstingen in het radiogebied, welke in de corona van de zon worden opgewekt na een grote zonnevlam. Daarna verbleef hij een jaar aan het Institute of Astronomy te Cambridge, U.K., om zich te bekwamen in de plasma-fysica van extragalactische radiobronnen. In 1976 trad hij wederom in dienst bij het Sterrekundig Instituut in Utrecht om onderzoek te verrichten in de extragalactische plasma-astrofysica.

en temperatuur voornamelijk oplichten in Hα, het ultraviolet of zachte röntgenstraling (figuur 2). Bij grotere zonnevlammen is steeds een dicht *filament* aanwezig, een reusachtige elektrische stroomdraad in de corona [1,2]. Bij de uitbarsting van de zonnevlam beweegt het filament zich met grote snelheid (500 km$\cdot$s^{-1}) omhoog. Hierbij kunnen zich schokken vormen, wordt de magnetische-veldstructuur herordend, worden versnelde deeltjes in magnetische lussen opgeslagen en ontsnappen energetische deeltjes langs 'open' veldlijnen naar de interplanetaire ruimte.

Als gevolg van deze verstoringen spelen zich in het magnetoplasma boven een activiteitscentrum allerlei plasmafysische processen af die wij gewaarworden aan een aantal radioverschijnselen [3]. Het is de kunst om plausibel te maken welk proces voor welk verschijnsel verantwoordelijk is en om de verschijnselen als diagnostisch middel te gebruiken voor de fysica van de actieve zon.

Plasma- en stralingsprocessen

Alvorens de belangrijkste solaire radiofenomenen te bespreken noemen wij enige processen in het coronale plasma die voor de interpretatie daarvan belangrijk zijn.

1. *Gyro-synchrotronstraling*. Een elektron dat zich in een magnetisch veld beweegt, ondervindt een lorentzkracht en zendt tengevolge hiervan elektromagnetische straling uit. Heeft het elektron een extreem relativistische energie, dan noemt men de uitgezonden continuümstraling synchrotronstraling. In het subrelativistische geval wordt de straling voornamelijk uitgezonden bij de (lagere) harmonischen van de cyclotronfrequentie en noemt men de straling gyrostraling. In het algemeen is in de corona de plasmafrequentie aanzienlijk groter dan de gyrofrequentie. Omdat nu de bij een zonnevlam versnelde deeltjes subrelativistisch zijn is gyrostraling veelal het belangrijkste stralingsmechanisme.

2. *Plasmastraling*. Indien de uitgezonden elektromagnetische straling niet meer kan worden opgevat als de somstraling van individuele deeltjes, spreekt men van plasmastraling. Het plasma speelt in dat geval een essentiële rol in de opwekking van de elektromagnetische straling. Veelal gebeurt dit als door het optreden van een instabiliteit het niveau van de altijd al wel aanwezige (thermische) plasmagolven aanmerkelijk versterkt wordt. Een deel van die plasma-energie kan door koppeling van de plasmagolven onderling of met andere golfsoorten in radiostraling worden omgezet die de corona verlaat.

3. *Instabiliteit*. Indien een (homogeen) plasma niet in thermisch evenwicht verkeert en voldoende ijl en heet is, kan het systeem relaxeren door de versterking van plasmagolven (die steeds in de ruis voorkomen). De kinetische energie van een bepaalde groep hoog-energetische deeltjes neemt daarbij af en komt ten goede aan elektromagnetische en bewegingsenergie van het hele plasma in de vorm van gol-

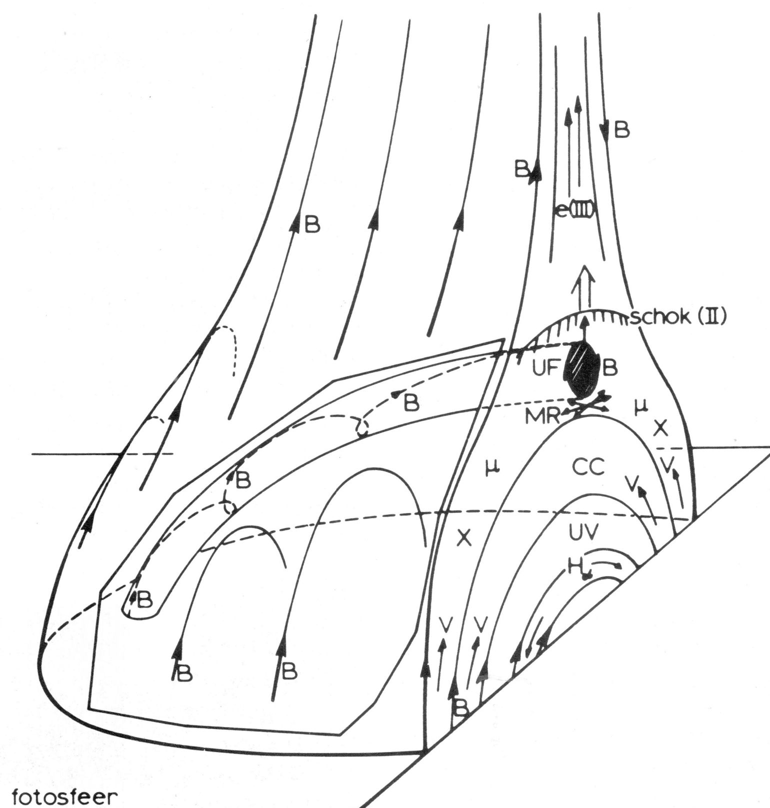


Fig. 2. Schematische voorstelling van de gebeurtenissen tijdens een zonnevlam in dwarsdoorsnede.
UF: uitbarstend filament
MR: reconnectie van magnetische velden en deeltjesversnelling
μ: microgolfstraling
CC: coronale condensatie ($T \simeq 3 \times 10^6$ K)
UV: ultraviolette emissie
Hα: lus met invallend koel materiaal ($T = 2 \times 10^4$ K)
X: röntgenstraling
V: 'verdamping' van dicht gas.
e: elektronen
B: magnetisch veld

ven. De aard van de opgewekte golven hangt af van de toestand van het verstoorde plasma. Twee belangrijke instabiliteiten zijn de zogenaamde verlieskegelinstabiliteit en de 'knobbel-in-de-staart' instabiliteit. Zo'n 'knobbel' geeft aanleiding tot plasma-Cerenkov-straling van Langmuir-golven. Dit zijn elektrostatische golven bij de plasmafrequentie van de elektronen. Op de verlieskegel komen wij nog terug.

4. *Omzetting van golven*. De bewegingsvergelijking van de plasmadeeltjes bevat de niet-lineaire convectieve term en de niet-lineaire lorentzkracht. Dat wil zeggen dat de respons van de deeltjesbeweging wordt bepaald door zowel de verstoorde beweging zelf als door het verstoorde veld. Bijgevolg kunnen de velden van meerdere plasmagolven samenwerken en nieuwe elektrische stromen (elektronenbewegingen) produceren die zelf weer als bron fungeren voor gelijksoortige of andere plasmagolven en voor (transversale) elektromagnetische straling die de (aardse) waarnemer kan bereiken. Is het golfniveau voldoende hoog, dan combineren de interagerende golven tot nieuwe golven op de som- en de verschilfrequenties.

Indien een hoogfrequente golf koppelt met een andere laagfrequente golf of een fluctuatie in de deeltjesverdeling, wordt straling gegenereerd bij nagenoeg de frequentie zelf van die (hoogfrequente) golf. Is dat een plasmagolf, dan wordt radiostraling als 'fundamental' uitgezonden. Treedt echter koppeling op tussen twee hoogfrequente golven onderling, dan kan combinatiestraling optreden bij de eerste harmonische frequentie ('harmonic').

De ruisstorm

De ruisstorm is het meest veelvuldig voorkomende fenomeen op metergolflengten (λ=1 à 5 m) en gaat altijd gepaard met het optreden van een zonnevlekkengroep. De straling is afkomstig van een brongebied op een hoogte van 0,2 à 0,6 $R_\odot$ (hoe korter de golflengte hoe lager; 1 $R_\odot$ is één zonstraal: 6,96 x 10^5 km) en met een uitgebreidheid van 2 à 4 boogminuten van de aarde af gezien. De ruisstorm bestaat uit een relatief langzaam variërend achtergrondscontinuüm met daarop gesuperponeerd kortstondige (0,3 à 0,6 sec.) smalbandige (3 à 6 MHz) stoten (figuur 4). De meeste stormen zijn sterk circulair gepolariseerd.

De stormstoten vertonen helderheidstemperaturen tot 10^{12} K, overeenkomend met een deeltjesenergie van 100 MeV. Zulke energierijke deeltjes worden slechts bij hoge uitzondering, door de krachtigste zonnevlammen, geproduceerd. Van de radio-uitbarstingen is de ruisstorm wel het minst begrepen. Wegens de hoge helderheidstemperatuur lijkt de verklaring te moeten worden gezocht in een coherent

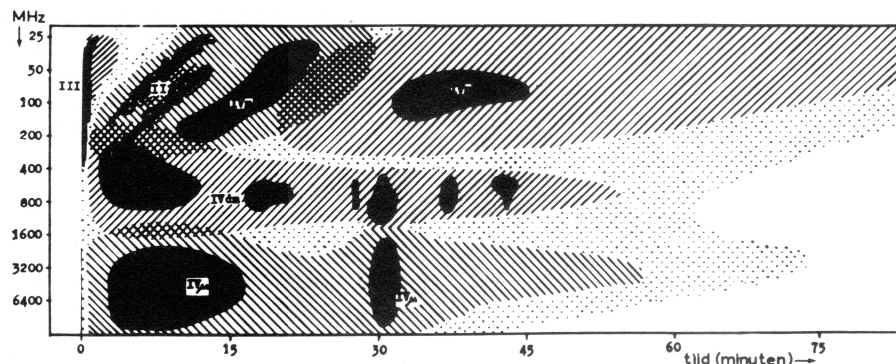


Fig. 3. Schematische weergave van de soorten radio-uitbarstingen tijdens en na een zonnevlam. Naast snel driftende uitbarstingen (type III) komen ook langzamere driften voor (type II) en continua zonder noemenswaardige frequentiedrift (type IV en μ).

stralingsmechanisme[4]. De stoten zijn veelal gerangschikt in fijne structuren die een zwakke frequentiedrift vertonen overeenkomend met een karakteristieke magneto-akoestische snelheid ter plaatse. Bijvoorbeeld kan men denken aan magnetische pulsen die plaatselijk de snelheidsverdeling van de elektronen instabiel maken en leiden tot de opwekking van plasmagolven op frequenties even boven de plasmafrequentie $f_p = (Ne^2/\epsilon_0 m)^{1/2}/2\pi$, f_p (Hz) = $8{,}9\sqrt{N}$ (N in m^{-3}, e is de lading, m de massa van een elektron en ϵ_0 de permittiviteit van vacuüm).
Het continuüm zou dan worden verklaard door een algemeen veld van plasmagolven. Deze zouden in stand worden gehouden door energierijke elektronen die in een magnetisch veld worden vastgehouden. De ruisstorm lijkt er op te wijzen dat er min of meer permanent elektronen worden versneld, misschien wel op aanmerkelijke hoogten (0,5 à 1 $R_\odot$).

Bundels deeltjes

Snel driftende ('Type III') uitbarstingen zijn kortstondige (1 à 2 s) stoten, meestal in groepen optredend, die in de loop van 1 à 2 seconden een groot frequentiegebied (van de orde van een octaaf of meer) in het meter- of decametergebied van hoog naar laag doorlopen (figuur 5). Veel type III-uitbarstingen gaan gepaard met een zonnevlam, maar vooral de decameter-type III's (die eigenlijk een categorie op zichzelf vormen) treden onafhankelijk van vlammen op.

Men vat deze uitbarstingen op als het effect van een snel langs een magnetische veldlijn naar buiten bewegende bundel elektronen. Deze bundel genereert Čerenkov-plasmastraling en de zo opgewekte Langmuir-golven geven op hun beurt aanleiding tot radio-emissie nabij de plasmafrequentie ('fundamental') en/of op de eerste harmonische daarvan. De meeste type III-uitbarstingen corresponderen vermoedelijk met de harmonische frequentie.

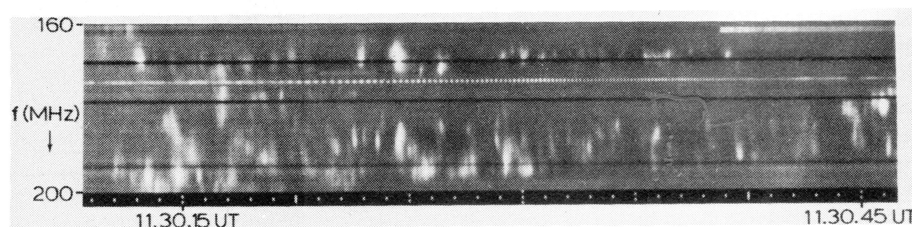
In de veronderstelling dat de emissie plaatsvindt op hetzij de plasmafrequentie of het dubbele daarvan, kan men uit een dichtheidsmodel van de corona de snelheid afleiden waarmee het verstorende agens naar buiten beweegt. Men vindt snelheden van de orde van $c/3$. (c is de lichtsnelheid.)

Satellietwaarnemingen hebben deze interpretatie bevestigd. De evolutie evenwel van de plasmagolven die bij de knobbel-in-de-staart-instabiliteit worden opgewekt is nog steeds niet begrepen. De laatste tijd is er van astronomische zijde dan ook een sterke belangstelling voor de zogenoemde modulatie van Langmuirgolven en solitonen.

Schokgolven

Bij grote zonnevlammen ziet men op radiogolflengten een zogenoemde type II-uitbarsting[3]. De radiostraling bestaat uit

Fig. 4. Radiospectrogram van een ruisstorm op 8 januari 1971 met type I-uitbarstingen die gerangschikt zijn in twee ketens met zwakke negatieve frequentiedrift. De opname is gemaakt met de 60-kanaals radiospectrograaf (thans te Dwingeloo). De zwarte kanalen zijn uitgevallen en het regelmatige onderbroken signaal is door mensen gemaakt.



twee parallelle banden die in frequentie een octaaf uiteenliggen (figuur 6). De emissie verschuift met de tijd van hoge naar lage frequenties. Uit interferometrische waarnemingen heeft men afgeleid dat het hier gaat om een verstoring die zich met een snelheid van meer dan 1000 $km \cdot s^{-1}$ door de corona naar buiten beweegt. Directe wit-licht-opnamen vanuit een satelliet laten zien dat binnen de radiostructuur de dichtheid verhoogd is. Deze soort radio-uitbarstingen is vermoedelijk van schokken afkomstig die zich met een snelheid van twee- tot driemaal de lokale Alfvén-snelheid voortplanten. (De Alfvén-snelheid is de snelheid waarmee geringe magnetische verstoringen zich langs het veld voortplanten.) In het (botsingsloze) schokfront verandert het magnetische veld aanzienlijk over een korte afstand.

Er zijn dus sterke elektrische stromen die vermoedelijk aanleiding geven tot de opwekking van elektronenplasmagolven (Langmuir-golven). Verstrooiing en onderlinge koppeling van deze plasmagolven leiden dan tot elektromagnetische straling bij de fundamentele en bij de eerste harmonische van de lokale plasmafrequentie. Dat deze twee mechanismen inderdaad in het spel zijn vindt een fraaie bevestiging in de harmonische structuur van de type II-uitbarsting.

Magnetische flessen

Continuüm ('Type IV') uitbarstingen zijn algemene verhogingen van het ruisniveau die zich over het grootste deel van het radiospectrum manifesteren. Vrijwel altijd is een type IV-uitbarsting het gevolg van het optreden van een grote zonnevlam, waarbij plotseling elektronen tot (sub)relativistische snelheden worden versneld. Elektronen die worden vastgehouden in een magnetische lus spiraliseren heen en weer tussen magnetische 'spiegels' en stralen daarbij (gyro-)synchrotronstraling uit. Maar tevens werken zij in op het plasma doordat zij dit instabiel maken voor plasmagolven. In het bijzonder speelt hierbij een rol de verlieskegelinstabiliteit.

Met verlieskegel wordt bedoeld dat de richtingsverdeling van de snelheden der snelle elektronen anisotroop is doordat snelle elektronen met bewegingsrichting binnen een zekere kegel met as evenwijdig aan het magnetisch veld ondervertegenwoordigd zijn (figuur 7). Zulke elektronen, met geringe spoedhoek, dringen namelijk door tot dichtere lagen waarin ze in korte tijd (ongeveer 10^{-4} s) hun energie door botsingen verliezen. Deze anisotrope verdeling van snelle elektronen is gesuperponeerd op een dichte isotrope thermische achtergrondverdeling die veel kouder is ($T \approx 2 \times 10^6$ K). Deze anisotropie leidt tot het ontstaan van plasmagolven op een ver-boven-thermisch energieniveau. Een kleine fractie van de plasma-energie lekt weg in de vorm van radiofrequente stralingsenergie.

In de verschillende spectrale gebieden (microgolf, decimeter, meter) heeft het type IV-continuüm van elkaar verschillende kenmerken (figuur 3) en de stralingsmechanismen zijn, mede vanwege de verschillende hoogten van oorsprong van de straling, niet noodzakelijk identiek. Op bijzonder karakteristieke wijze manifesteert de type IV-uitbarsting zich in het decimetergebied door een hoge helderheidstemperatuur (10^{11} K), een smalle bandbreedte ($<$ 100 MHz) en door de merkwaardige modulaties die dikwijls in het continuüm optreden (figuur 8). Wederom wijzen deze waarnemingen op een collectieve aard van het emissiemechanisme. Op radiospectrografische opnamen met goed tijd- en frequentiescheidend vermogen en goede contrastwerking leveren die een intrigerend schouwspel op. In hoofdzaak zijn er drie van zulke modulaties die wij nu zullen bespreken.

1. *Zebrapatronen*. In het frequentietijdvlak zijn dit in een quasi-periodiek patroon gerangschikte vermeerderingen en verminderingen van de intensiteit. De typische afstand tussen aangrenzende 'ribbels' is 3 à 8 MHz. Een zebrasysteem verschuift dikwijls in frequentie, waarbij telkens veranderingen in dat verloop optreden.

Dit verschijnsel kan worden geduid als een verlieskegelinstabiliteit die leidt tot de opwekking van plasmagolven op de bovenste hybride frequentie $f_{uh} = (f_p^2 + f_g^2)^{1/2}$ op plaatsen waar de verhouding van de frequentie f_{uh} tot de gyrofrequentie f_g heeltallig is ($f_g = (eB)/(2\pi m)$, f_g (Hz) $= 2{,}8 \times 10^{10}\, B$(T), B is de magnetische fluxdichtheid). Een fractie van de energie van deze golven wordt omgezet in elektromagnetische straling op nagenoeg dezelfde frequentie.

Indien de variatie van f_{uh} langs de verschillende 'resonante' oppervlakken voor de verschillende heeltallige verhoudingen f_{uh}/f_g voldoende klein is ($\Delta f_{uh} < f_g$), dan zullen op verschillende frequenties emissiebanden optreden (figuur 8). Aangezien de lokatie van de resonante oppervlakken erg gevoelig is voor de precieze veld- en dichtheidsstructuur in het brongebied, zal een geringe magnetohydrodynamische relaxatie leiden tot een merkbare verschuiving van de 'resonante' oppervlakken naar andere frequenties; vandaar de 'drift' in de zebrapatronen[5]. Indien daarentegen het brongebied niet homogeen genoeg is, zullen de bovenste hybride frequenties langs de verschillende oppervlakken elkaar overlappen en ziet men niets anders dan een continuüm.

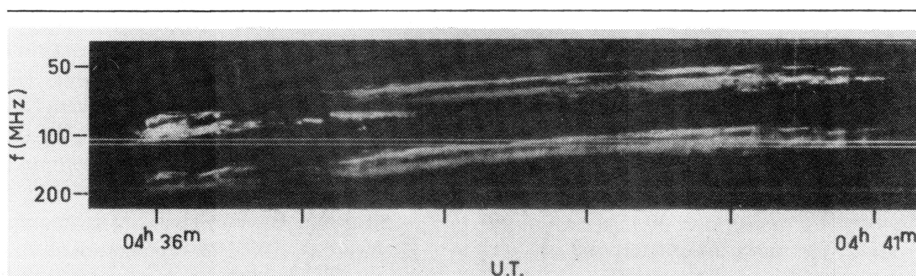


Fig. 6. Radiospectrogram van een type II-uitbarsting op 9 october 1969 opgenomen met de radiospectrograaf te Culgoora. (G.A. Dulk, Proc. Astron. Soc. Australia 1 (1970) 308)

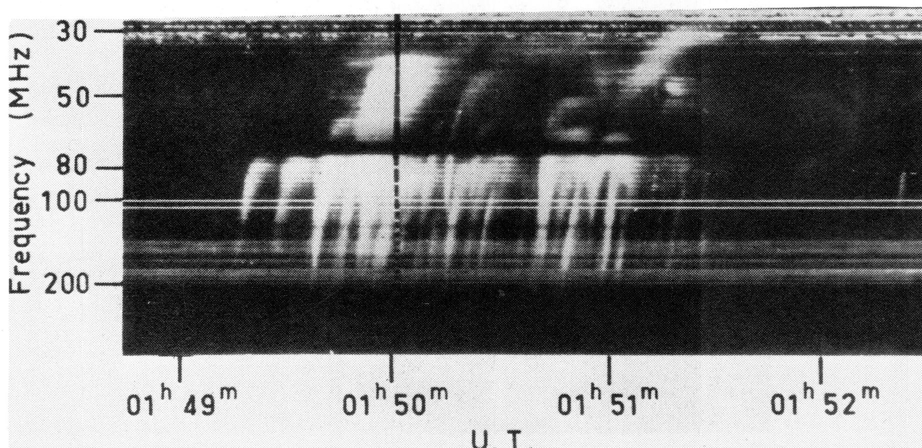


Fig. 5. Radiospectrogram van een groep type III-uitbarstingen op 25 februari 1971 opgenomen met de radiospectrograaf te Culgoora. (I.D. Palmer, R.P. Lin, Proc. Astron. Soc. Australia 2 (1972) 101)

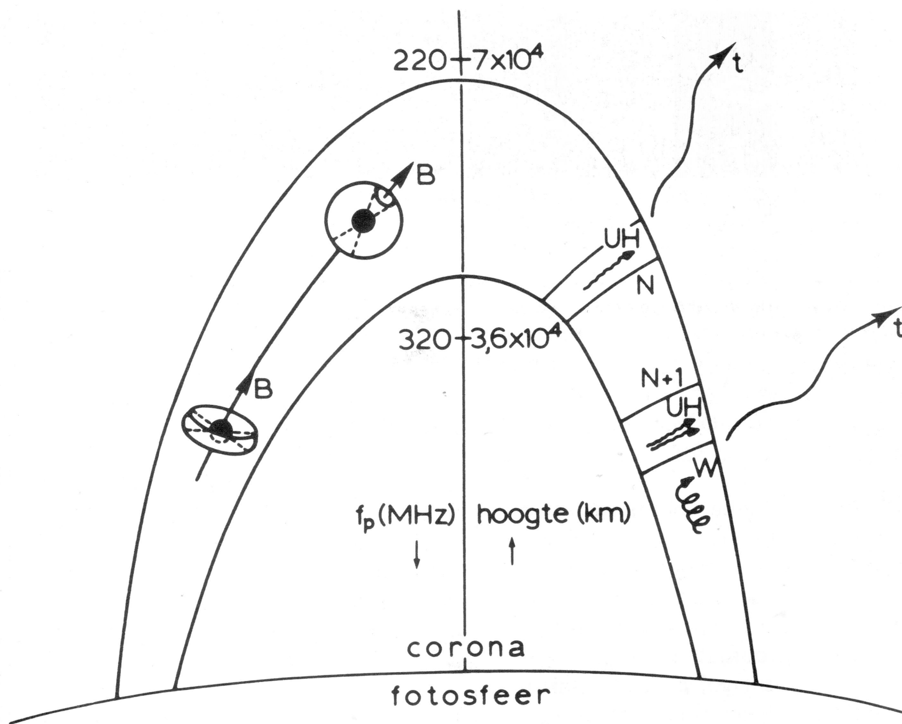


Fig. 7. Schets van het brongebied in doorsnee van een type IV-continuüm met fijne structuren. De snelheidsverdeling van de deeltjes is aangegeven op twee plaatsen in de magnetische lus. De verschillende golfsoorten zijn aangeduid met:
uh: bovenste hybride.
w: whistler
t: ontsnappende elektromagnetische straling.
Verder zijn twee opeenvolgende 'resonante' oppervlakken aangegeven, waarop f_{uh}/f_g heeltallig is (gelijk aan N en respectievelijk N+1).

2. *Fiber uitbarstingen*. Deze modulatie ziet eruit als een geïsoleerde ribbel in het tijd-frequentievlak met een 'emissie'kant en een aangrenzende 'absorptie'kant aan laagfrequente zijde. Daarbij ziet het ernaar uit dat een hoeveelheid energie is overgeheveld van de ene frequentie naar een naburige ($\Delta f \approx 3$ MHz) hogere frequentie. De fiber vertoont een driftsnelheid die inligt tussen die van type III- en van type II-uitbarstingen.

Dit verschijnsel kan worden geduid als het tweede effect van de verlieskegelinstabiliteit waarbij zgn. 'whistlers' worden opgewekt.

Een whistler is een elektromagnetische golf op een frequentie f_w lager dan de gyrofrequentie, die een voorkeursrichting van voortplanting langs het magnetische veld heeft. Whistlers zijn welbekend van de karakteristieke audiofrequente fluittoontjes die worden geproduceerd door bliksems en die langs aardmagnetische veldlijnen heen en weer reizen. Whistlers kunnen niet uit de corona ontsnappen, en hun aanwezigheid in de corona kan dan ook niet direct worden aangetoond.

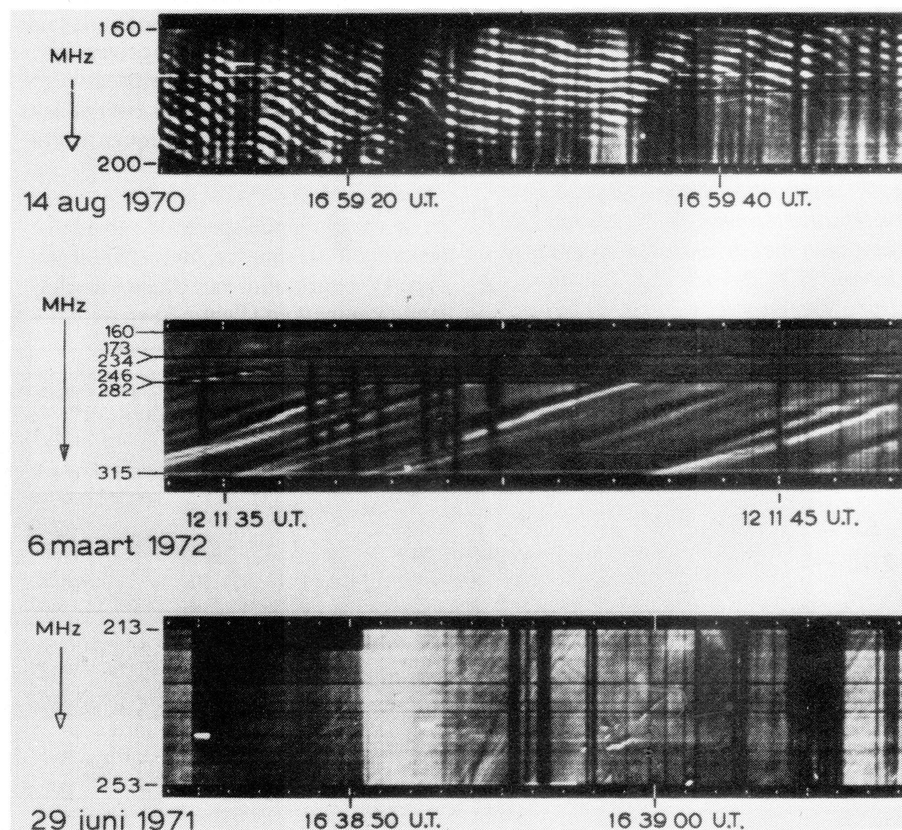
In de corona zouden whistlers bij voorkeur worden opgewekt nabij de 'voetpunten' van magnetische lussen, waar de anisotropie van de elektronenrichtingsverdeling maximaal is. Als golfpakketjes reizen de whistlers langs het veld omhoog. Waar ze de veel hoger frequente ($f = f_{uh}$) plasmagolven ontmoeten, zullen ze daarmee koppelen en elektromagnetische straling produceren op de frequentie $f_{uh} + f_w$.

(N.B. Bij $f_{uh} - f_w$ kan elektromagnetische straling zich niet voortplanten in het plasma.) Plasmagolven die combineren met een whistler zijn daarmee verloren voor het opwekken van elektromagnetische straling (door strooiing aan fluctuaties) op hun eigen f_{uh} – frequentie; vandaar wellicht de met de emissie gepaard gaande intensiteitsreductie (figuur 8).

Het frequentieverloop van de fiber zou een afspiegeling zijn van de naar buiten gerichte verplaatsing van het whistler-pakket (met een snelheid van 21 – 28 maal de Alfvén-snelheid), terwijl de waargenomen richelafstand met de whistlerfrequentie overeenkomt (0,1 – 0,4 f_g). Uit beide gegevens kan men onafhankelijk de plaatselijke veldsterkte in de bron bepalen. Inderdaad blijken deze waarden met elkaar in overeenstemming[5].

3. *Plotselinge reducties, versnelling in pulsen*. De intensiteit van de type IV-continuümstraling vertoont af en toe een serie reducties die vrijwel simultaan plaatsvinden over een frequentiebereik van de orde van een half octaaf à een octaaf. Even plotseling als de vermindering van intensiteit geschiedt, komt na ca. 0,3 s de intensiteit

Fig. 8. Voorbeelden van fijne structuren in type IV-continua opgenomen met de radiospectrograaf te Dwingeloo:
bovenaan: zebrapatronen;
midden: fibers;
en onder: plotselinge reducties.



weer terug op het uitgangsniveau. Binnen een serie zijn de afzonderlijke reducties meestal onregelmatig in de tijd verdeeld, maar soms herhalen ze zich quasiperiodiek.

Ter verklaring denkt men aan de mogelijkheid dat de verlieskegelinstabiliteit wordt opgeheven door de plotselinge injectie van snelle deeltjes (elektronen) die, met een bewegingsrichting ruwweg langs het magnetisch veld, de verlieskegel tijdelijk opheffen[7]. Aangezien een isotrope maar verder willekeurige deeltjesverdeling stabiel is, zodat geen plasmagolven worden opgewekt, zullen de bestaande instabiliteiten onmiddellijk worden onderdrukt en de reeds aanwezige golven worden gedempt.

Met deze interpretatie verplaatst zich het probleem naar de mogelijkheid van een mechanisme waarbij versnelling pulsgewijs optreedt. Er zijn twee mogelijkheden geopperd.
Volgens het ene mechanisme zouden ionakoestische golven, opgewekt door elektrische stromen in het vlamgebied, aanleiding geven tot Langmuir-golven[8]. Energie van die golven komt ten goede aan elektronen die door Landau-demping worden versneld. Deze energie-uitwisseling treedt op tussen golven en deeltjes waarvan de snelheid in de richting van de golfvector gelijk is aan de fasesnelheid van de betreffende golf. Een en ander wordt beschreven door een stelsel van gekoppelde niet-lineaire vergelijkingen (genoemd naar Lotka en Volterra) die slechts periodieke oplossingen toelaten. Oorspronkelijk waren die vergelijkingen opgesteld ter verklaring van periodieke fluctuaties in populaties van roof- en prooidieren.

Volgens het andere mechanisme zou versnelling in zonnevlammen tot stand komen door middel van geïnduceerde elektrische velden, analoog aan de in het laboratorium geproduceerde pulsen van 'runaway' deeltjes[9].

Een aanwijzing voor het optreden van pulsgewijze versnelling wordt ook geleverd door de zeer kortstondige ($\lesssim$ 0,02 s) stralingspieken die zijn gevonden in registraties van microgolfuitbarstingen met hoog tijdscheidend vermogen[10].

Besluit
De diagnostiek van de radiovlam zelf, in het microgolfgebied, staat in feite nog in de kinderschoenen. Voor het komende Solar Maximum Mission-jaar staat een unieke combinatie van satelliet- en grondwaarnemingen op stapel om de zonnevlam te bestuderen. De Nederlandse sterrenkundige gemeenschap zal daaraan een tweeledige bijdrage leveren: enerzijds zal de radiostraling van vlammen worden bestudeerd met de Westerbork Synthese Radio Telescoop (voor positie-informatie) en de radiospectrograaf te Dwingeloo (voor fijne structuren in frequentie en tijd). Anderzijds zal de röntgenstraling gelijktijdig worden waargenomen met de Hard X-ray Imaging Spectrometer in de SMM-satelliet.

Referenties
1. W. van Tend: *Magnetohydrodynamic and thermal processes in solar flare energy build-up and release*. Proefschrift, RU-Utrecht, 1979.
2. M. Kuperus, H. Rosenberg, NTvN **39** (1973) 130.
3. J.P. Wild, S.F. Smerd, Annual Rev. Astron. Astrophys. **10** (1972) 159.
4. A. Kattenberg, R.M. van Hees, J. van Nieuwkoop, in: *Radio Physics of the Sun*. M. Kundu and T. Gergeley (eds.) IAU Symp. No. 86 (1979), in press.
5. J. Kuijpers, Astron. Astrophys. **40** (1975) 405.
6. J. Kuijpers, Solar Phys. **44** (1975) 173.
7. A.O. Benz, J. Kuijpers, Solar Phys. **46** (1976) 275.
8. J. Kuijpers, Astron. Astrophys. **69** (1978) L9.
9. C.S. Liu, Y.C. Mok, K. Papadopoulos, F. Engelmann, M. Bornatici, Phys. Rev. Letters **39** (1977) 701.
10. C. Slottje, Nature **257** (1978) 520.

Boekrecensies

Thomas S. Kuhn,
Black-Body Theory and the Quantum Discontinuity 1894-1912.
Clarendon Press: Exford University Press, 1978.
XIII, 356 blz., £ 12.50, gebonden.

In het algemeen laten fysische leerboeken de quantummechanica beginnen met Planck's lezing op 14 december 1900 over de theorie van de energieverdelingsfunctie bij 'zwarte stralers'. Ook vooraanstaande wetenschapshistorici als Kangro, Klein, Hermann en Jammer nemen dit standpunt in. Zij interpreteren de introductie van de hypothese $\epsilon = h\nu$ als een quantisatie van de energie E van een oscillator: $E = 0, h\nu, 2h\nu, \ldots$
Kuhn – onder wetenschapsfilosofen en wetenschapshistorici geen onbekende – verdedigt in zijn boek de uitdagende stelling, dat deze visie berust op het onjuist lezen van Planck's artikel. Planck wil namelijk Boltzmann's waarschijnlijkheidsdefinitie van de entropie toepassen op het probleem van de zwarte straling. Hij telt daartoe het aantal oscillatoren met een energie ergens tussen $nh\nu$ en $(n + 1)h\nu$, waarbij Planck volgens Kuhn uitdrukkelijk niet bedoelt – maar wel zegt – dat de energie uitsluitend de randwaarden van het interval zou kunnen aannemen. In essentie komt Kuhn's stelling erop neer, dat Planck een zuiver *klassieke, continue* energieverandering voor ogen stond, terwijl iedereen tot nu toe gedacht heeft, dat Planck de energie van een oscillator slechts *discrete* waarden toekende en dat hij daarmee de eerste *quantum-voorwaarde* formuleerde.

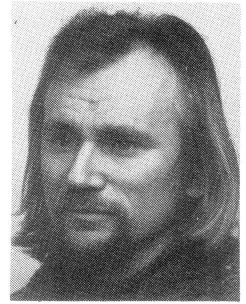
Zo'n standpunt schept verplichtingen. Het strekt tot lof van Kuhn, dat hij op nauwgezette wijze zijn stelling tracht te bewijzen. In het eerste deel van zijn boek gaat hij in op de weg waarlangs Planck kwam tot zijn energieverdelingsfunctie. Daarbij bespreekt hij onder meer de ideeen van Boltzmann en Planck over irreversibiliteit, moleculaire wanorde en het gebruik van combinatoriek voor de definitie van de entropie. Op zeer gedetailleerde wijze behandelt Kuhn de verschillende fases in Planck's theorie. Hij tracht aan te tonen dat zowel Planck als zijn lezers in de periode 1900-1906 zijn theorie niet als een breuk met de klassieke fysica ervoeren. In het tweede deel van het boek beschrijft Kuhn de groei van het besef dat Planck's theorie een essentiële afwijking van gevestigde denkbeelden impliceert. Twee artikelen van Einstein en Ehrenfest uit 1906 vormden daartoe de eerste aanzetten. Pas nadat Lorentz in 1908 het probleem zeer scherp had gesteld, begon een wat grotere groep fysici zich van de moeilijkheden bewust te worden. Tegen 1910 was een aantal prominente fysici, waaronder Planck zelf, ervan overtuigd dat op een of andere wijze een discontinuïteit ingevoerd diende te worden. In een afsluitend derde gedeelte beschrijft Kuhn Planck's zogenaamde 'tweede theorie' van 1911-12. Deze theorie zag men tot nu toe als het krampachtig vasthouden van een koppige oude man aan de klassieke fysica. Kuhn draait ook deze waardering 180° om: naar zijn mening is dit juist Planck's eerste (gepubliceerde) erkenning van het niet-klassieke karakter van zijn theorie.

Bij de beoordeling van het boek lijkt het volgende me van belang. Kuhn doet alsof er slechts twee, zeer scherp te bepalen, posities mogelijk zijn: Planck voerde op 14 december 1900 discontinue energie op atomaire schaal in, of hij deed dat niet. Het eerste standpunt werd tot nu toe ingenomen, Kuhn verdedigt het tweede. Deze polariserende manier van voorstellen bevordert de duidelijkheid van zijn eigen stelling en maakt een vruchtbaar debat mogelijk. Het moeilijk te interpreteren historisch materiaal geeft evenwel misschien een mogelijkheid de twee posities een zodanige 'lijnbreedte' te geven dat ze elkaar overlappen.

C. Hakfoort

Het Einstein-observatorium voor hoge-energieastrofysica

J. Heise



John Heise (37) studeerde af in 1969 in de theoretische sterrenkunde aan de RUU en is sedertdien werkzaam in het Laboratorium voor Ruimte-onderzoek van het Sterrekundig Instituut te Utrecht.
Hij specialiseerde zich in de röntgenastronomie en was betrokken bij de voorbereiding, begeleiding en verwerking van meetresultaten verkregen met de röntgendetectoren aan boord van de astronomische Nederlandse satelliet ANS.
Sinds kort werkt hij aan de interpretatie van waarnemingen met het Utrechtse tralie-experiment in het Einstein-observatorium.

Op 13 november 1978 werd in de Verenigde Staten de tweede satelliet in de serie High Energy Astrophysical Observatories, HEAO-B, gelanceerd, die ter gelegenheid van het honderdste geboortejaar van Albert Einstein, het Einstein-observatorium werd gedoopt. De röntgentelescoop aan boord heeft een ruimtelijk scheidend vermogen en een gevoeligheid, die een verbetering betekent van een factor 100, respectievelijk 1000 ten opzichte van het tot dan toe gebruikelijke. Hiermee is HEAO-B het eerste volwassen röntgen-observatorium, waarvan de resultaten een kwaliteit bezitten vergelijkbaar met die van de huidige grootste optische en radiotelescopen.
De Nederlandse betrokkenheid is veelzijdig. Het Laboratorium voor Ruimteonderzoek te Utrecht heeft een Objective Grating Spectrometer vervaardigd, bestaande uit twee soorten transmissietralies van resp. 500 en 1000 lijnen/mm. De Universiteit van Leiden heeft een samenwerkingsprogramma voor het waarnemen van dezelfde velden aan de hemel met de synthese-radiotelescoop in Westerbork, en verscheidene individuele astrofysici hebben gastwaarnemingsprogramma's voor waarnemingen met het Einstein-observatorium ingediend. De eerste resultaten zijn beschreven in een speciaal aan het Einstein-observatorium gewijd nummer van Astrophysical Journal Letters (november 1979).
In dit artikel beoog ik een korte samenvatting te geven van de status van de röntgenastronomie tot nu toe om daarna aan de hand van de illustraties een indruk te geven van de potentiële mogelijkheden van dit röntgenobservatorium.

Röntgenastronomie
Onder het röntgenvenster in het elektromagnetische spectrum wordt meestal het fotonenergiegebied van 0,1 ....100 keV verstaan. Hierin onderscheidt men de m.b.v. scherende inval te focusseren zachte röntgenstraling van 0,1 ....3 keV en de alleen met mechanische middelen te collimeren harde röntgenstraling van 3....100 keV.
De kosmos blijkt een grote verscheidenheid van röntgenbronnen te kennen. Vóór de lancering van het Einstein-observatorium waren er zo'n 300 bronnen bekend plus een ruimtelijk onopgeloste 'diffuse' component in de röntgenemissie. De röntgenstraling van bronnen van galactische oorsprong blijken, voorzover geïdentificeerd, veelal samen te hangen met de laatste stadia van sterevolutie. Jonge supernovarestanten emitteren over een uitgestrekt gebied van de orde van een lichtjaar synchrotronstraling afkomstig van elektronen van ongeveer 10^{11} eV in een magneetveld van ongeveer 10^{-3} Gauss, die beide gegenereerd worden door een snel roterende gemagnetiseerde neutronenster, het geïmplodeerde restant van de supernova-explosie. De roterende neutronenster zelf is als radio-röntgen- en gammastralingspulsar waarneembaar. De uiteindelijke energiebron is de rotatie-energie van de neutronenster (voorbeelden: de Krab-pulsar en de Vela-pulsar).

Andere röntgenbronnen blijken samen te hangen met neutronensterren in nauwe dubbelstersystemen. Hier is de energiebron van gravitationele oorsprong: materie valt van de begeleidende ster (bijvoorbeeld door een sterwind) in het gravitatieveld van de neutronenster, vormt ten gevolge van het specifieke impulsmoment een dunne schijf in Kepler-banen rond de neutronenster. Door de differentiële rotatie en viscositeit in deze schijf diffundeert de materie naar binnen tot een afstand waarop magnetische krachten van het magneetveld van de neutronenster domineren. Op kleinere afstand wordt de materie langs de veldlijnen geleid tot de magnetische polen. Dit proces wordt aangeduid als accretie, de schijf heet accretieschijf. Bij accretie op een neutronenster komt er in de gravitatiepotentiaal een hoeveelheid energie vrij overeenkomend met ongeveer 10% van de rustmassa-energie. Het proces is zelfregulerend in die zin, dat de stralingsdruk bij te grote accretie verdere accretie verhindert. De maximaal waargenomen absolute flux van 10^{38} erg/s reflecteert deze grens. Er is nu een twintigtal röntgendubbelsterren bekend, waaronder Hercules X-1. In deze laatste bron is de grootte van het magnetisch veld aan het oppervlak van de neutronenster aan de hand van het optreden van cyclotronabsorptielijnen bij 60 keV en zijn eerste harmonische bij 120 keV bepaald op ongeveer 10^{13} Gauss.

Een derde klasse bronnen zijn de z.g. galactische builbronnen in en in de nabijheid van het galactisch centrum. Zij behoren zowel relatief als absoluut tot de helderste röntgenbronnen, maar vertonen geen enkele aanduiding voor het lidmaatschap van een dubbelstersysteem. De door de Astronomische Nederlandse Satelliet ANS ontdekte z.g. röntgenbursters, uitbarstingen die binnen één seconde plaatsvinden en tot minuten kunnen duren, behoren tot deze klasse. De precieze aard van deze bronnen is nog niet bekend, hoewel er sterke aanwijzingen zijn dat ook de röntgenbursters neutronensterren zijn.
Extragalactische röntgenemissie is gedetecteerd in elk type sterrenstelsel, maar is zeer groot voor actieve stelsels zoals Seyfert-stelsels, radiostelsels en quasars. De precieze aard van de emissie is niet geheel bekend, maar er is een toenemend aantal aanwijzingen dat de emissie voor een belangrijk deel is geconcentreerd in een uiterst klein gebied nabij het centrum van deze stelsels. Daarnaast zijn clusters van melkwegstelsels ook, maar nu ruimtelijk uitgebreide, röntgenbronnen. Mede door het detecteren van Fe-lijnen nabij 6 keV, leidt men af dat de emissie hier remstraling is door een ijl, heet (10^8 K) interclustergas.
De oorsprong van de omnidirectionele-diffusieröntgenstraling was tot voor kort nog duister. Geïnterpreteerd als waarlijk diffuse straling afkomstig van een heet gas (5×10^8 K) tussen de clusters van de melkwegstelsels heeft het belangrijke kosmologische consequenties. Een belangrijke parameter in kosmologische modellen is namelijk de materiedichtheid in het heelal. Bij een gemiddelde dichtheid kleiner dan een kritieke waarde heeft het heelal een oneindig volume ('het heelal is open') en expandeert eeuwig. Bij een dichtheid groter dan de kritieke waarde is het volume eindig ('het heelal is geslo-

ten') en de huidige expansie verandert t.z.t. in een contractie. De dichtheid van de zichtbare materie is slechts ongeveer 20% van de kritieke dichtheid.
Eventueel overal aanwezige ijle koele materie zou door absorptie gemakkelijk gedetecteerd kunnen worden en blijkt niet aanwezig. De overblijvende mogelijkheid is dus hete materie. Een dichtheid van ongeveer 1 atoom per kubieke meter is voldoende om het heelal te 'sluiten'.
Echter het verrassend grote aantal bronnen waargenomen met het Einstein-observatorium doet vermoeden dat een belangrijke bijdrage wordt geleverd door ruimtelijk onopgeloste puntbronnen.
Het vraagstuk is dus weer open, het heelal waarschijnlijk ook.

Het Einstein-röntgenobservatorium
Het 6 meter lange in doorsnede metende Einstein-observatorium maakt gebruik van de volledige reflectie van zachte röntgenstraling (4 ... 120 Å) bij scherende inval onder een hoek van ongeveer 1°. Een combinatie van achtereenvolgens hyperbolisch en parabolisch gevormde spiegels geeft een afbeelding van de röntgenhemel met een gezichtsveld van 75 x 75 boogminuten in het brandvlak. In dit brandvlak kunnen verschillende detectoren worden geplaatst. De belangrijkste twee zijn een plaatsgevoelige proportionele teller (de IPC, Imaging Proportional Counter) met een resulterend ruimtelijk scheidend vermogen van 1 boogminuut en een channel array plate (HRI, High Resolution Images), waarbij het ruimtelijk scheidend vermogen door de optische eigenschappen van de spiegel is beperkt tot 2 boogseconden. In de stralenbundel kan een door het Laboratorium voor Ruimteonderzoek te Utrecht vervaardigde transmissietralie worden geplaatst, waardoor een spectrum verkregen kan worden met een spectraal scheidend vermogen van 0,5 à 1 Å tussen 5 en 120 Å. Zie voor een volledige beschrijving van het experiment Giacconni, referentie 2.

Supernova-restanten
Een dankbaar object van studie voor het Einstein-observatorium is de röntgenemissie van supernova-restanten, vanwege hun ruimtelijke uitgebreidheid (van de orde van een boogminuut). De 1000 jaar oude Krab-nevel (figuur 1) ziet er in röntgenstraling ongeveer zo uit als in gepolariseerd licht. Duidelijk blijkt de centrale pulsar die pulseert met een periode van 33 milliseconde en de verhoogde emissie ten noorden van de pulsar, die overeenkomt met een bekend nevelig actief gebied uit optische opnamen. Ook in de 10 000 jaar oude Vela-supernova-restant is de centrale pulsar, die met een periode van 89 ms al bekend was uit radio-, optische en gammastralingsmetingen, nu ook in röntgenemissie waargenomen. Wonderlijk genoeg blijkt in enkele jongere supernova-restanten (enkele honderden jaren), zoals Cassiopeia A (figuur 2 uit Murray[4]) en in Tycho's supernova-restant niets van een centraal object. Deze afwezigheid impliceert een bovengrens van $1{,}5 \times 10^6$ K van een mogelijke stellair restant, een factor 5 lager dan de verwachte temperatuur van zo'n recent gevormde neutronenster. Als niet in (bijna) alle supernova-explosies neutronensterren worden gevormd dan is dit een extra moeilijkheid in het rijmen van de grote neutronensterdichtheid (afgeleid uit het grote aantal radiopulsars) en de toch al relatief lage frequentie van waargenomen supernova-explosies (2 tot 4 per eeuw) in ons melkwegstelsel.

Een andere mogelijkheid is dat er wel een centrale neutronenster aanwezig is, maar dat de oppervlaktetemperatuur veel lager is dan verwacht. Daar de koeling van een neutronenster de eerste 1000 jaar ná de (verondersteld hete) vorming volledig wordt bepaald door neutrinoverliezen uit de kern, kan dit betekenen dat andere tot nu toe niet beschouwde processen een rol spelen. Dit zou bijvoorbeeld het geval zijn als de gevormde neutronenster een kern bezit, die bestaat uit pionen.

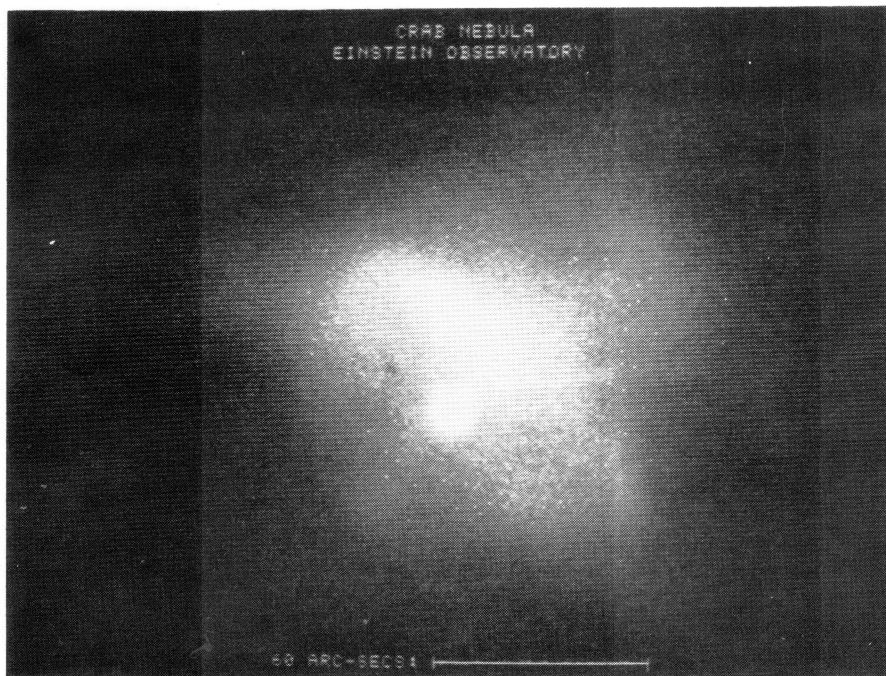


Fig. 1. Opname van de Krab-nevel met het Einstein-observatorium.

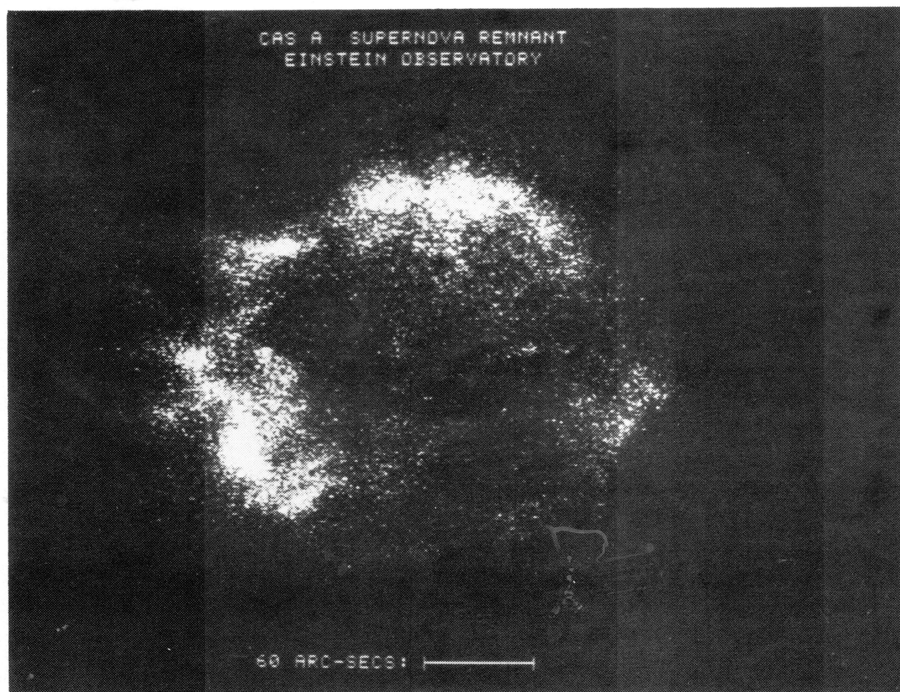


Fig. 2. Het Einstein-observatorium: de supernovarestant Cas A in röntgenemissie. Uit: Murray et al. (1979). Courtesy Astrophysical Journal Letters.

Stellaire bronnen

Een van de verrassendste eerste resultaten van het Einstein-observatorium is het grote aantal röntgenbronnen dat gevonden is. Al uit waarnemingen met de Astronomische Nederlandse Satelliet ANS bleek dat er behalve 'klassieke' galactische bronnen, die samenhangen met neutronensterren, ook andere typen röntgenbronnen bestaan. Deze stellaire bronnen, zoals Sirius, Capella, vlamsterren e.d. hebben een absolute helderheid, die 10^4 tot 10^6 keer kleiner zijn dan de röntgendubbelstersystemen. De door ANS en door sondeerraketwaarnemingen gedetecteerde stellaire bronnen blijken nu de uiterste top van de ijsberg. Praktisch iedere ster emitteert een zekere fractie van zijn optische lichtkracht in röntgenstraling. Deze fractie varieert van ongeveer 10^{-3} voor hete sterren (oppervlaktetemperatuur $>$ 10 000 K) via een minimum van 10^{-6} voor sterren iets heter dan de zon, naar een fractie van ongeveer 0,1 voor de koelere sterren (N.B. vergelijk dit met een factor 10^4 voor de verhouding röntgenemissie tot optische emissie in een bron als Scorpius X-1). De aard van de stellaire röntgenemissie is waarschijnlijk remstraling uit een optisch dun plasma met een temperatuur van 10^6 tot 10^7 K, dat zich bevindt in een ijle corona rond de ster. Van koele sterren, zoals de zon, is het bekend dat een corona kan ontstaan ten gevolge van het feit dat de buitenste lagen in de ster in convectief evenwicht zijn.
In hete sterren is de theorie nog niet duidelijk, maar hangt het bestaan van een corona waarschijnlijk mede samen met de grote stralingsdruk en de daardoor lage netto versnelling aan het oppervlak van de ster.

Andromeda-nevel

De bestudering van röntgenbronnen in onze naaste buur, het spiraalstelsel M31 (de Andromeda-nevel) biedt zekere voordelen ten opzichte van de bronnen in ons eigen melkwegstelsel. De afstanden zijn bekend en daarom is bijvoorbeeld de absolute helderheidsverdeling nauwkeuriger te bepalen. De spiraalstructuur en verdere morfologie is duidelijker zichtbaar dan in ons stelsel zodat bijvoorbeeld bronposities eenduidiger zijn te interpreteren.
Vóór de waarnemingen van het Einstein-observatorium was bekend dat de Andromeda-nevel een zwakke, nog juist te detecteren röntgenbron was. Figuur 3 toont de resultaten (Van Speybroeck[5]) met de positiegevoelige proportionele teller (ruimtelijke resolutie 1′). De nevel bestaat uit een groot aantal puntbronnen. Het diffuse centrale, in dit plaatje ruimtelijke onopgeloste deel blijkt ook uit een groot aantal (45) puntbronnen te bestaan, zoals de opname in figuur 4 met de High Resolution Imager (resolutie 2″) aantoont. In totaal zijn 69 röntgenbronnen te onderscheiden. Na vergelijking met een optische opname is de spiraalstructuur in de posities van röntgenbronnen duidelijk waarneembaar. In totaal is zo'n 60% van de geïdentificeerde bronnen geassocieerd met z.g. populatie I-objecten (jonge sterassociaties, gebieden met interstellaire stof, gebieden van geïoniseerd waterstof enz). De resterende 40% behoort tot de galactische builbronnen (galactic 'bulge' sources) nabij het centrum van M31. Hiervan is zo'n 10% geïdentificeerd met bolvormige sterhopen. Naast de overeenkomsten, zijn natuurlijk ook de verschillen met ons eigen stelsel interessant. Opvallend is de bron, die binnen de nauwkeurigheid van 2″ in de kern van M31 wordt waargenomen. Met een helderheid van 10^{38} erg/s is deze als röntgenbron niet ongewoon sterk, maar is wel een factor 1000 sterker dan de bronnen binnen 0°.5 van het centrum van ons eigen stelsel. De radio-emissie van de kern van M31 is slechts 6% (op 1415 MHz) van die van Sgr A, de radiokern van de melkweg. Dit suggereert dat de röntgenemissie in M31 van thermische aard is, mogelijk accretie van gaswolken op een zwaar zwart gat (10^6 zonmassa's). De constantheid van de emissie is een verdere ondersteuning van dit model.

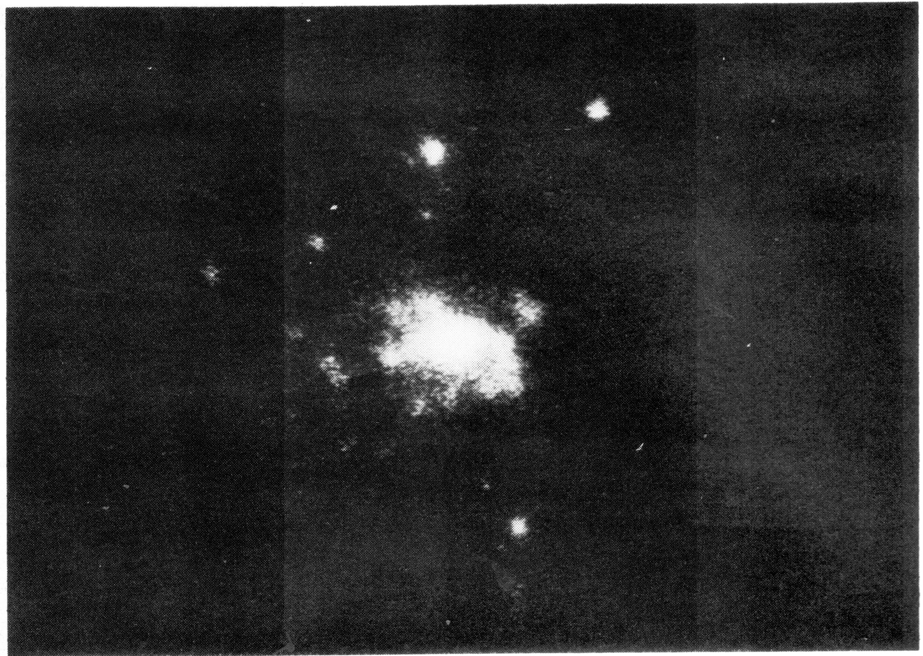


Fig. 3. Het Einstein-observatorium: de spiraalnevel M31 in röntgenemissie (resolutie 1′). Uit: Van Speybroeck et al. (1979). Courtesy Astrophysical Journal Letters.

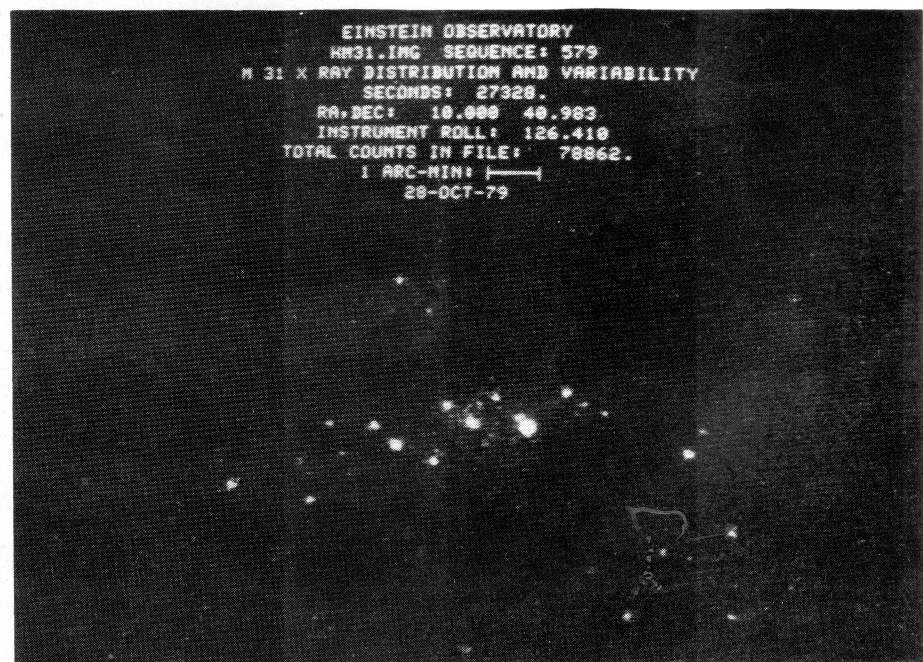


Fig. 4. Het Einstein-observatorium: het centrum van de spiraalnevel M31 in röntgenemissie met een resolutie van 2″. Uit: Van Speybroeck et al. (1979). Courtesy Astrophysical Journal Letters.

Clusters van melkwegstelsels

Röntgenwaarnemingen zijn van vitaal belang voor het begrijpen van de structuur

en evolutie van clusters van melkwegstelsels en is nu eerst goed mogelijk met het Einstein-observatorium. Een zeer groot deel van de waarnemingsprogramma's is dan ook gericht op extragalactische röntgenastronomie. Het lijkt erop dat een cluster begint als een grote wolk van stelsels, die geleidelijk instort onder invloed van zijn eigen gravitatieveld om tenslotte een dynamisch evenwicht te bereiken wanneer een dichte kern en een uitgestrekte halo van melkwegstelsels is gevormd. Men kan de snelheidsdispersie van de melkwegstelsels in een cluster gebruiken als maat voor de gravitatiepotentiaal. De Virgo-cluster heeft een lage dispersie en is een voorbeeld van een cluster in een vroeg evolutiestadium. Figuur 5 toont enkele röntgen-isofoten (Forman[1]) gesuperponeerd op een optische opname van een deel van de groep. Bijna ieder stelsel blijkt een ruimtelijk uitgebreide diffuse bron van röntgenstraling te zijn veroorzaakt door remstraling van een ijl (4×10^{-3} deeltjes per cm^3) heet gas met een temperatuur van 10^7 K. Blijkbaar is dit gas gebonden door de gravitatiepotentiaal van het stelsel. Dit geheel is gesuperponeerd op een diffuse harde röntgenbron (met een uitgebreidheid overeenkomend met de gehele Virgo-groep van stelsels, veel groter dan de opname in figuur 3). De harde röntgenmissie is afkomstig van remstraling van een nog ijler (5×10^{-4} deeltjes cm^{-3}) gas met een temperatuur van 10^8 K dat wordt gebonden door de gravitatiepotentiaal van de Virgo-cluster als geheel. Forman et al.[1]) interpreteren de lokale 10^7 K component als gas dat wordt 'opgeveegd' door de eigen beweging van de stelsels ten opzichte van het hete 10^8 K-intraclustergas. Een vergelijkende studie (Jones[3]) van de röntgenemissie van verschillende rijke, nabije clusters van melkwegstelsels bevestigt dit beeld. De clusters waarin de röntgenemissie is opgebroken in klonten rond individuele stelsels vertonen een lagere temperatuur en lage snelheidsdispersie, terwijl de clusters met hoge snelheidsdispersie een gladde, centraal gecondenseerde röntgenintensiteitsverdeling vertonen waarin de temperatuur veel hoger is. In dit laatste geval wordt de verdeling van gas meer bepaald door de gravitatiepotentiaal van de cluster als geheel dan door die van de individuele stelsels. Deze clusters hebben ook een kleinere fractie aan spiraalstelsels, daar het interstellaire gas en stof in deze stelsels is opgeveegd door het dichtere en hetere intraclustergas.

De transmissietralies

Het gebruik van transmissietralies is een nieuwe en veelbelovende techniek in de röntgenspectroscopie ontwikkeld in het Laboratorium voor Ruimteonderzoek te Utrecht. De spectrale resolutie van proportionele tellers voor zachtere röntgenstraling met een golflengte groter dan bijvoorbeeld 20 Å is buitengewoon slecht ($\Delta\lambda/\lambda \sim 1$), terwijl voor de tralies $\Delta\lambda/\lambda$ gelijk is aan ongeveer 0,5 Å/λ. Ten tijde van het prepareren van de tralies voor het Einsteinobservatorium was de technologie echter nog niet zodanig dat tralies van voldoende dikte gemaakt konden worden, zodat de efficiencies nog erg laag zijn (ongeveer 20% van het theoretisch mogelijke). Deze onvolkomenheid is nu overwonnen: de tralies die worden voorbereid voor een volgend röntgenobservatorium, de Europese EXOSAT, voldoen volledig aan de theoretische verwachting. Figuur 6 toont een opname van de heldere röntgenbron Cygnus X-1 met de Objective Grating Spectrometer (1000 lijnen/mm).

De centrale piek is de nulde orde van deze intense, door de spiegel- en steunroosterrespons schijnbaar uitgebreid lijkende puntvormige bron, met aan weerszijden de eerste orde van 5 tot 25 Å. Zwak zichtbaar is ook het tweede-orde-beeld, dat slechts 1% van de eerste orde bedraagt. Emissie van röntgenstraling met een golflengte langer dan 25 Å is onderdrukt door fotoionisatie van waterstof en helium in het interstellaire medium tussen bron en waarnemer. De vorm van het continuüm-

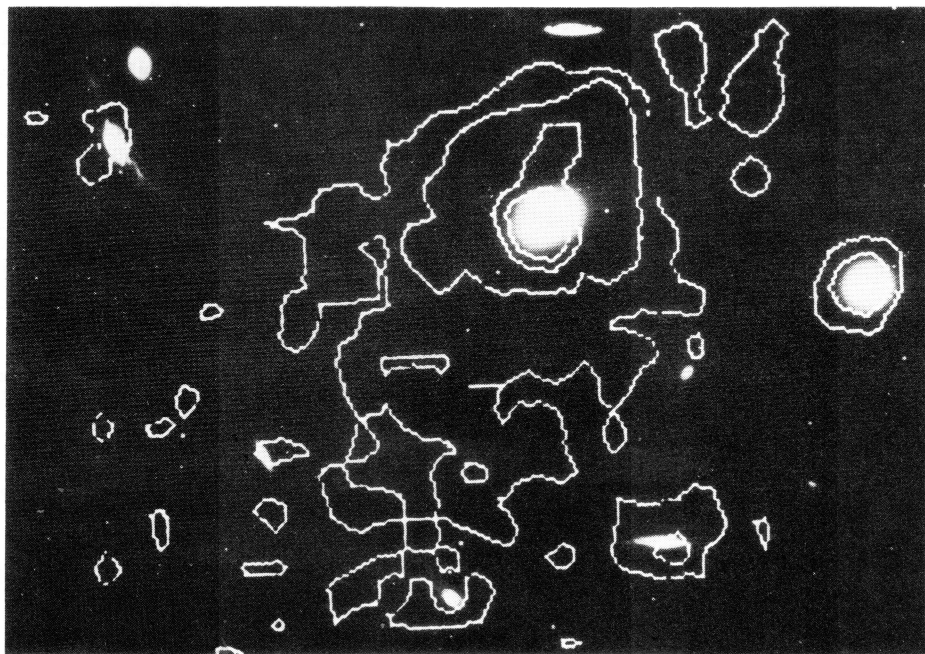


Fig. 5. Het Einstein-observatorium: het centrale deel van de Virgo-cluster met daarop gesuperponeerd isofoten van het röntgenbeeld. Uit: Forman et al. (1979). Courtesy Astrophysical Journal Letters.

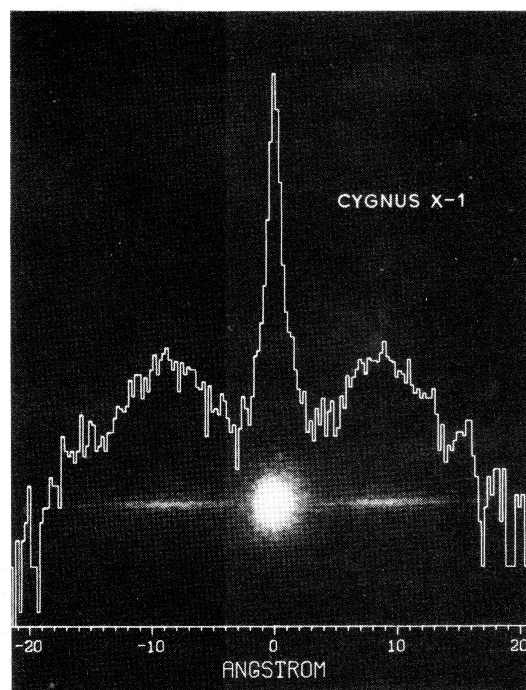


Fig. 6. Het spectrum van de röntgenbron Cygnus X-1 opgenomen met de Objective Grating Spectrometer van het Einstein-observatorium.

spectrum samen met de op deze foto niet duidelijk zichtbare absorptie-K-kant van zuurstof bij 23,3 Å levert informatie op over de chemische samenstelling van het interstellaire medium.

Conclusie

Deze zes plaatjes geven slechts een globale indruk van de potentiële mogelijkheden van het Einstein-observatorium.

Vele reeds gepubliceerde resultaten (Quasars, Deep Sky Survey, radiostelsel Centaurus A, de discrete-bronbijdrage tot de extragalactische achtergrond enz.) zijn onbesproken gebleven. Nog meer resultaten liggen klaar voor nadere analyse. De satelliet zal werkzaam zijn tot het voorjaar van 1981. De röntgenastronomie, ontstaan door de toevallige ontdekking van de bron Scorpius X-1 in 1962, is daarmee de kinderjaren ontgroeid en behoort met de optische, de UV-, en de radioastronomie tot een volwaardige tak van de observationele sterrenkunde.

Referenties

1. W. Forman et al., Ap. J. Letters **234** (1979) L 27.
2. R. Giaconni et al., Ap. J. **230** (1979) 540.
3. C. Jones et al., Ap. J. Letters **234** (1979) L 21.
4. S.S. Murray et al., Ap. J. Letters **234** (1979) L 69.
5. L. van Speybroeck et al., Ap. J. Letters **234** (1979) L 45.

De dubbelsterpulsar en de algemene relativiteitstheorie

H.F. Henrichs en E.P.J. van den Heuvel

Naast de algemene relativiteitstheorie van Einstein is er een aantal 'concurrerende' relativistische theorieën die de zwaartekrachtwisselwerking beschrijven. Elk van deze theorieën doet bepaalde voorspellingen die, in principe, aan de waarnemingen getoetst kunnen worden. Tot voor kort waren de experimentele onzekerheden in de voorspelde effecten nog zodanig groot dat op grond hiervan nog vele van deze zwaartekrachttheorieën naast elkaar mogelijk bleven.

Een recente ontdekking in de astronomie heeft er echter toe geleid dat een aantal 'relativistische' effecten met verrassende nauwkeurigheid gemeten kan worden. Dit artikel poogt een nadere achtergrond van deze ontdekking, een radiopulsar in een dubbelstersysteem, te geven.

Alle aan deze dubbelsterpulsar gemeten verschijnselen blijken met Einsteins theorie in overeenstemming te zijn, doch zullen in de nabije toekomst waarschijnlijk de meeste andere theorieën doen afvallen. Speciaal het optreden van gravitatiestraling trekt hierbij de aandacht. Men mag verwachten dat in de komende jaren de meetnauwkeurigheid nog aanzienlijk zal worden verhoogd.

Newtons gravitatiewet, geïnspireerd op de drie wetten van Kepler, opende de mogelijkheid om de banen van hemellichamen nauwkeurig te beschrijven en te voorspellen. Sinds Newtons tijd is het planetenstelsel steeds bij uitstek het terrein geweest voor het toetsen van gravitatietheorieën. De reden hiervoor is vooral dat de zon het enige object in onze buurt is dat een zeer sterk zwaartekrachtveld bezit, waarin de bewegingen van testdeeltjes (de planeten) nauwkeurig kunnen worden bestudeerd.

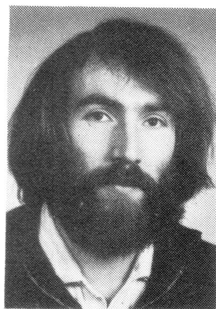
Ook toen Einstein in 1915 de algemene relativiteitstheorie had opgesteld, bleek al spoedig dat een aantal van de door deze theorie voorspelde correcties op Newtons theorie (zoals de periheliumdraaiing van de mercuriusbaan en de afbuiging van het licht langs de zonrand) in het zonnestelsel kan worden gemeten. Toch is er nog een aantal door de algemene relativiteitstheorie voorspelde verschijnselen die niet in het zonnestelsel zijn te verifiëren omdat het zwaartekrachtveld van de zon hiertoe niet voldoende sterk is. Dit geldt in het bijzonder voor tweede- en hogere-orde-effecten, zoals het optreden van gravitatiestraling. Het blijkt dat een aantal van deze hogere-orde-effecten in de zogenaamde dubbelsterpulsar wèl gemeten kan worden. Deze pulsar werd in 1974 ontdekt door Hulse en Taylor[1] van de University of Massachusetts te Amherst. Ze ontdekten dat de pulsar, die de naam PSR 1913+16 draagt (de getallen geven ruwweg de coördinaten aan de hemel aan), zich in een nauwe, zeer excentrische (e = 0,61) baan beweegt om een (onzichtbare) begeleider. De omloopperiode is minder dan 8 uur, en de afstand tussen de twee sterren is van de orde van de straal van de zon.

Pulsars zijn zeer nauwkeurige klokken en het feit dat zo'n precisieklok zich in een sterk gravitatieveld beweegt, opende de mogelijkheid om in een tamelijk kort tijdbestek relativistische effecten te meten. Om een indruk te geven van de sterkte van de algemeen-relativistische effecten in deze dubbelster: een bekend resultaat van de algemene relativiteitstheorie is de extra periheliumdraaiing van de elliptische baan van mercurius rond de zon. De lange as van de mercuriusbaan verdraait ongeveer 43″ per *eeuw* en een nauwkeurige meting van dit effect kost vele decennia. In de dubbelsterpulsar bedraagt daarentegen de waargenomen relativistische draaiing van de lange as meer dan 4° per *jaar*, en deze kon reeds binnen twee jaar na de ontdekking van de pulsar nauwkeurig worden gemeten. Dit is uiteraard het gevolg van het veel sterkere gravitatieveld in deze dubbelster vergeleken met dat in ons zonnestelsel. Een maat voor de sterkte van het veld is GM/c^2r, waarbij G de gravitatieconstante is, c de lichtsnelheid en r de afstand tot het middelpunt van het object met massa M. Het blijkt dat deze grootheid voor PSR 1913+16 ruim 100 maal groter is dan voor mercurius.

De draaiing van de lange as is een eerste-orde-correctie op de theorie van Newton. De relativistische correcties op de baanbeweging treden omloop na omloop op en leiden na vele omlopen op cumulatieve wijze tot grote afwijkingen ten opzichte van de door Newtonse gravitatie voorspelde posities. Vier jaar na de ontdekking van de pulsar waren er zoveel meer en nauwkeuriger waarnemingen bekend, dat voor het eerst inderdaad een tweede-orde-afwijking kon worden gemeten: de omloopperiode blijkt korter te worden. Taylor, Fowler en McCulloch[2] interpreteerden deze metingen aan de hand van Einsteins theorie. In dit artikel wordt nader ingegaan op deze interpretatie.

Radiopulsars

Radiopulsars[3] zijn radiobronnen waarvan de emissie een duidelijk gepulst karakter heeft (zie figuur 1). De eerste radiopulsar werd in 1967 ontdekt door een groep in Cambridge, Engeland, geleid door A. Hewish, die hiervoor in 1974 de Nobelprijs voor Natuurkunde ontving.

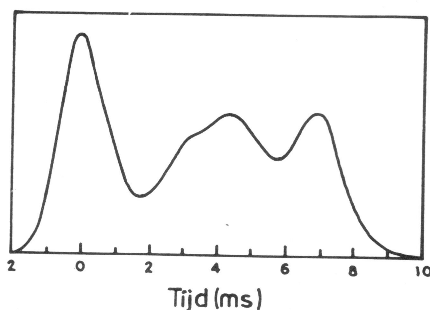


Drs. H.F. Henrichs (30) studeerde aan de Universiteit van Amsterdam waar hij de doctoraalexamens experimentele natuurkunde (1975, Zeeman Laboratorium) en sterrenkunde (1976) aflegde. Sindsdien bereidt hij zich voor op zijn proefschrift (aanvankelijk in ZWO-verband) onder leiding van prof. dr. E.P.J. van den Heuvel op het Sterrenkundig Instituut van de Universiteit van Amsterdam. Onderzoek wordt verricht aan fysische processen die samenhangen met het massatransport in röntgendubbelsterren, waarbij de aandacht geconcentreerd ligt rond röntgenpulsars.

Sindsdien zijn er meer dan 300 pulsars gevonden. Pulsars blijken een zeer stabiele pulsperiode te bezitten met een waarde tussen de 0,03 en 5 seconden. De stabiliteit van deze 'klokken' is van de orde $\delta P/P = 10^{-12}$. De algemeen aanvaarde verklaring voor de stabiliteit van deze klokken is dat de *rotatie* van een zwaar object de pulsperiode bepaalt.

Bij pulsars blijkt het object een *neutronenster* te zijn. De massa is vergelijkbaar met die van de zon, terwijl de diameter ongeveer 20 km bedraagt. De materie van een neutronenster bestaat vrijwel geheel uit gedegenereerde neutronen. De gemiddelde dichtheid is dus ongeveer die van kernmaterie: ruwweg 10^{17} kg$\cdot$m^{-3}. Bovendien blijken neutronensterren een extreem sterk magnetisch veld te bezitten met grotendeels een dipoolkarakter. Aan het oppervlak bedraagt dit veld 10^8 T. De radioemissie blijkt opgewekt te worden in gebieden in de buurt van de magnetische polen. Het pulserende karakter ontstaat vanwege het niet samenvallen van de rotatie-as en de magnetische dipool-as (zie figuur 2). De fysica van de magnetosfeer van een pulsar is zeer gecompliceerd en de *precieze* locatie van de opwekking der radio-emissie is nog steeds niet duidelijk. In elk geval liggen de bronnen ervan niet verder dan enkele duizenden kilometers van het steroppervlak. De uitgezonden radiostraling is slechts een kleine fractie van de totaal door de neutronenster uitgezonden energieflux. Het grootste deel van deze flux wordt geëmitteerd in de vorm van laagfrequente magnetische dipoolstraling opgewekt door de voortplanting van het ronddraaiende magnetische veld in de ruimte.
De energiebron van een pulsar is de rotatie van de neutronenster: de emissie van de magnetische dipoolstraling gaat ten koste van de rotatie-energie. We zien dan ook dat praktisch alle radiopulsars in de loop der tijd steeds langzamer gaan draaien en dus een steeds langere pulsperiode krijgen.

Fig. 1. Intensiteit van de radiopulsar PSR 1913+16 als functie van de tijd. Elke pulsar heeft zijn eigen specifieke pulsvorm.



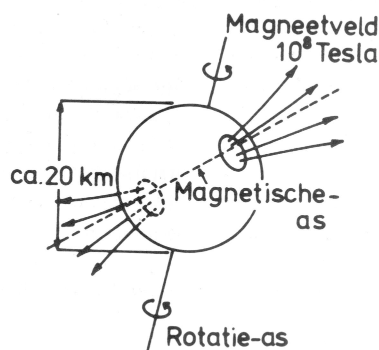
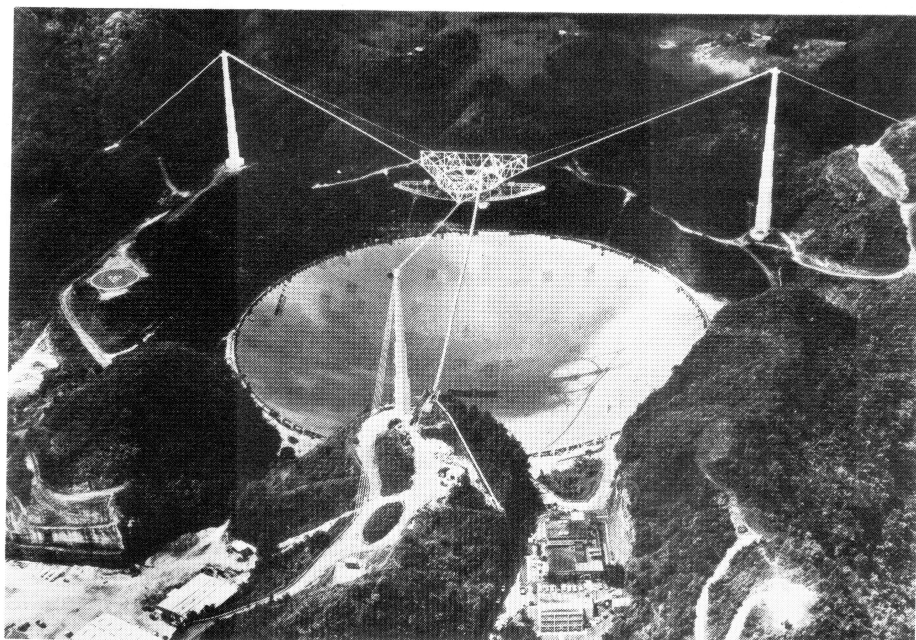


Fig. 2. Model voor een radiopulsar. De magnetische dipool-as valt niet samen met de rotatie-as. De kegelvormige radiobundels (niet getekend) zijn verbonden met de magnetische polen. Alleen waarnemers binnen het bereik van de kegelbundels zullen de gepulste radiostraling zien.

Fig. 3. Het Arecibo Observatorium in Puerto Rico. Met zijn 305 m diameter is deze (vast opgestelde) schotel de grootste radiotelescoop ter wereld. De natuurlijke vorm van het dal wordt als ondersteuning gebruikt. Bijna 40000 aluminium panelen vormen een sferisch oppervlak dat minder dan 3 mm afwijkt van de ideale vorm. De panelen zijn geperforeerd opdat de vegetatie eronder in leven blijft en daardoor de bodem stabiel houdt. De ontvanger in het brandpunt kan (zeer nauwkeurig) verplaatst worden zodat een gebied van 40° diameter aan de hemel bestudeerd kan worden.





Prof. dr. E.P.J. van den Heuvel (39) is hoogleraar sterrenkunde aan de Universiteit van Amsterdam. Hij studeerde van 1957 tot 1962 natuur- en sterrenkunde te Utrecht, waar hij in 1968 promoveerde op een studie over de rotatie van sterren. Hij werkte aan de RU-Utrecht, de Vrije Universiteit van Brussel (waar hij sinds 1970 Buitengewoon Hoogleraar is), de University of California te Santa Cruz en het Institute for Advanced Study te Princeton. Sinds 1974 is hij werkzaam te Amsterdam. Zijn onderzoeksterrein ligt op het gebied van de sterevolutie en de hoge-energie-astrofysica.

De dubbelsterpulsar

PSR 1913+16, door Hulse en Taylor in

juli 1974 met de reuzentelescoop van Arecibo in Puerto Rico[4] (zie figuur 3) ontdekt, heeft een periode van slechts 0,059 s, de op één na kortste periode van alle bekende radiopulsars. De metingen te Arecibo werden uitgevoerd bij 430 MHz en 1410 MHz (respectievelijk 70 en 21 cm). De gebruikte ontvanger detecteert de radiostraling in 32 aan elkaar grenzende frequentiebanden, elk met een bandbreedte van 250 kHz.

De keuze van deze ontvangsttechniek is bepaald door (afgezien van de pulsduur zelf) het feit dat het pulsarsignaal dispersie ondergaat in het interstellaire medium, als gevolg van de daar aanwezige vrije elektronen. Deze dispersie maakt dat de puls bij een hogere frequentie eerder arriveert dan bij een lagere. Door nu de ontvangen pulssignalen in de respectieve kanalen naar toenemende frequentie op de juiste wijze te vertragen, kan men voor de interstellaire dispersie corrigeren, en de aankomsttijden bij verschillende frequenties laten samenvallen, waarna de signalen opgeteld kunnen worden. Zodoende wordt gecompenseerd voor het verlies aan gevoeligheid dat anders door de interstellaire dispersie zou worden veroorzaakt. Deze dispersie hangt af van de afstand van de pulsar. Dat voor iedere pulsar de ontvanger aldus op specifieke wijze moet worden geoptimaliseerd is een bijkomstige moeilijkheid bij het opsporen van vooral zwakke pulsars.

Reeds een paar uur na de ontdekking van PSR 1913+16 bleek dat de periode niet constant is, maar op systematische wijze varieert tussen 0,058967 en 0,059045 s (zie figuur 4). De interpretatie was onmiddellijk dat deze pulsar zich in een baan rond een ander object beweegt. Dat de aankomsttijden erg slecht op een sinuskromme passen maakte duidelijk dat het om een zeer excentrische baan gaat. De complete afwezigheid van eclipsen wijst erop dat de onzichtbare begeleider kleiner moet zijn dan een ster die in het stadium van waterstoffusie verkeert (een zogenaamde hoofdreeksster, zoals de zon).

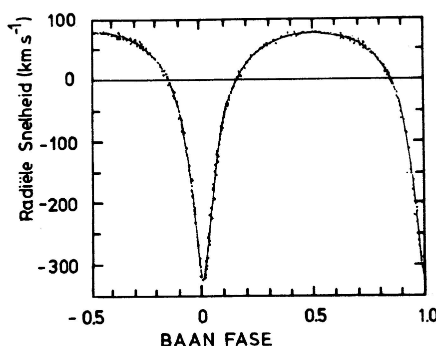


Fig. 4. De aankomsttijden (vertaald in radiale snelheden) van de dubbelsterpulsar als functie van de baanfase.

Meer conclusies waren na een maand niet gerechtvaardigd.

Na anderhalf jaar, in het begin van 1976, was het voor het eerst mogelijk om de verandering in de lengte van het periastron met redelijke nauwkeurigheid te meten (zie tabel I). Figuur 5 illustreert de definities van de diverse baanelementen.

De analyse van de aankomsttijden der pulsen

In de loop der jaren heeft men de aankomsttijden van de pulsen kunnen meten met een relatieve nauwkeurigheid van 50 μs. Van de functie die de aankomsttijden van de pulsen bij de aarde in de loop van de tijd weergeeft, verwacht men – op theoretische gronden – dat zij vastgelegd is door een 13-tal parameters. Deze zijn:

- De coördinaten α en δ van de pulsar aan de hemel. Het dopplereffect, als gevolg van de baanbeweging van de aarde rond de zon en de rotatie van de aarde om haar as, veroorzaakt veranderingen in de aankomsttijden, welke in de loop van het jaar variëren. De wijze van deze variatie hangt van α en δ af.
- De pulsperiode P en diens afgeleide naar de tijd, $\dot{P}$.
- De vijf keplerbaan-parameters (in figuur 5 is hun betekenis weergegeven). Deze veronderstellen newtonse mechanica en bepalen de hoofdmoot van de periodieke dopplerverschuivingen die het pulssignaal ondergaat. Tenslotte:
- Vier parameters van relativistische oorsprong, die de correcties op de newtonse baanbeweging weergeven. (Het meenemen van nog meer relativistische parameters is niet nodig daar deze in alle theorieën ordes van grootte kleiner zijn dan deze vier.)

Verder moet men nog een (bekende) algemeen-relativistische klok-correctie toepassen om in rekening te brengen dat het op aarde gevoelde zwaartekrachtveld van de zon in de loop van het jaar varieert.

Elk der 13 parameters beïnvloedt de aankomsttijden van de pulsen op zijn eigen wijze en, door de waarden van deze parameters te variëren, bepaalt men (met een kleinste-kwadratenmethode) die combinatie van parameterwaarden waarvoor een optimale aanpassing bij de waargenomen aankomsttijden wordt verkregen. In tabel I zijn de aldus verkregen waarden van de 13 parameters weergegeven met hun bijbehorende meetonzekerheden zoals bepaald door Taylor en medewerkers. De met deze 13 waarden verkregen functie, die de aankomsttijden der pulsen voorspelt, is zo goed dat men hiermee de pulsaankomsttijden 8 maanden van tevoren voorspellen kan met een nauwkeurigheid beter dan 100 microseconden!

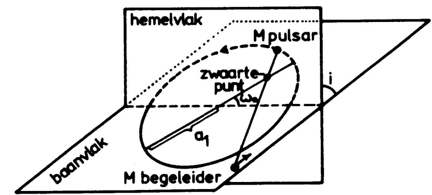


Fig. 5. De baan van de dubbelsterpulsar in de ruimte. De symbolen stemmen overeen met die in tabel I. De lengte van het periastron is ω_0. De halve lange as van de elliptische baan is a_1.

De relativistische parameters

De vier relativistische parameters zijn:

- $\dot{\omega}_{AR}$, de algemeen-relativistische hoeksnelheid van periastrondraaiing;
- γ, welke een maat is voor de (variabele) vertraging die de puls ondergaat in het dubbelstersysteem als gevolg van de gravitatieroodverschuiving gecombineerd met het transversale dopplereffect.
- $\sin i$, waarbij i de inclinatiehoek van het baanvlak is ten opzichte van de hemelbol (zie figuur 5). Deze beïnvloedt de vertraging van het pulssignaal in het zwaartekrachtveld van de begeleider.
- $\dot{P}_b$, de verandering van de dubbelsterperiode (als gevolg van emissie van gravitatiestraling).

Deze vier parameters zijn thans alle met een 95% of hogere betrouwbaarheid be-

Tabel I. *De 13 parameters voor de dubbelsterpulsar PSR 1913+16.*

Rechte klimming (1950,0)	α	$= 19^h 13^m 12\overset{s}{.}474(4)$
Declinatie	δ	$= 16°01'08\overset{''}{.}02(6)$
Pulsperiode	P	$= 0{,}059029995269(2)$ s
Tijdsafgeleide van periode	P	$= (8{,}64 \pm 0{,}02) \times 10^{-18}$
Geprojecteerde halve lange as	$a_1 \sin i$	$= 2{,}3424(7)$ lichtseconde
Excentriciteit	e	$= 0{,}617155(7)$
Baanperiode	P_b	$= 27906{,}98172(5)$ s
Lengte van het periastron	ω_0	$= 178\overset{\circ}{.}864(2)$
Tijd van periastronpassage	T_0	$=$ JD 2442321,433206(1)
Transversaal dopplereffect en gravitationele roodverschuiving	λ	$= 0{,}0047(7)$ s
Inclinatiehoek	$\sin i$	$= 0{,}81(16)$
Tijdsafgeleide van baanperiode	$\dot{P}_b$	$= (-3{,}2 \pm 0{,}6) \times 10^{-12}$
Snelheid van periastrondraaiing	$\dot{\omega}$	$= 4\overset{\circ}{.}226(2)$ jaar^{-1}

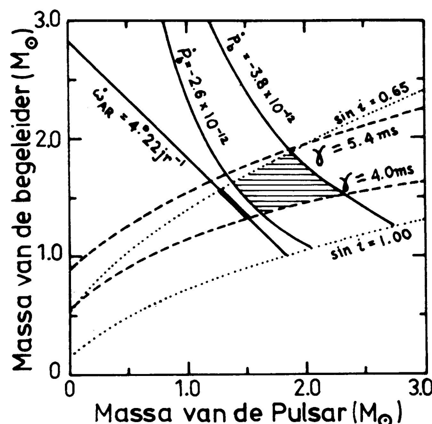
γ	$= 0{,}002951\,(M_p)\,(M_p + 2\,M_b)\,(M_p + M_b)^{-4/3}$ s
$\sin i$	$= 0{,}5083\,(M_b)^{-1}\,(M_p + M_b)^{2/3}$
$\dot{P}_b$	$= 1{,}70 \times 10^{-12}\,(M_p M_b)\,(M_p + M_b)^{-1/3}$
$\dot{\omega}_{AR}$	$= 2\overset{\circ}{.}11\,(M_p + M_b)^{2/3}$ jaar

Tabel II. *De massa-afhankelijkheid van de relativistische parameters. (alle massa's in zonne-eenheden)*

paald. Elk van deze vier parameters blijkt een functie te zijn van de onbekende massa's M_p en M_b van de pulsar en zijn begeleider (en van de waargenomen keplerbaanparameters, die zeer nauwkeurig bekend zijn). De precieze aard van deze vier functionele verbanden hangt nog af van de gebruikte gravitatietheorie. Voor het geval van de algemene relativiteitstheorie zijn deze vier functies in tabel II weergegeven. Meting van de vier parameters levert vier vergelijkingen op waaruit men de onbekende massa's M_p en M_b zal moeten bepalen.

Daar er meer vergelijkingen dan onbekenden zijn is het de vraag of de M_p- en M_b-waarden die door oplossing van de verschillende vergelijkingen verkregen worden binnen de grenzen van de meetnauwkeurigheid met elkaar in overeenstemming zijn. Indien dit het geval is, is de toegepaste theorie bruikbaar. Het blijkt dat dit voor de algemene relativiteitstheorie inderdaad het geval is. Figuur 6 toont de gebieden in het (M_p, M_b) vlak welke door

Fig. 6. De vier gemeten relativistische parameters leveren vier vergelijkingen op voor de massa's M_p en M_b van de pulsar en zijn begeleider (zie tabel II). Van deze vier is $\dot{\omega}_{AR}$ met zeer grote nauwkeurigheid gemeten; daar deze alleen van de som $M = M_p + M_b$ afhangt, levert hij de teruglopende lijn op. Voor de andere 3 (hogere-orde relativistische) parameters is het 95%-betrouwbaarheidsinterval aangegeven, hetgeen gezamenlijk het gearceerde gebiedje levert. Men mag stellen dat deze twee gebieden niet significant verschillen, gezien de onzekere grenzen van het gearceerde gebied. Samen levert dit $M_p = 1{,}39 \pm 0{,}15\ M_\odot$ en $M_b = 1{,}44 \pm 0{,}15\ M_\odot$.



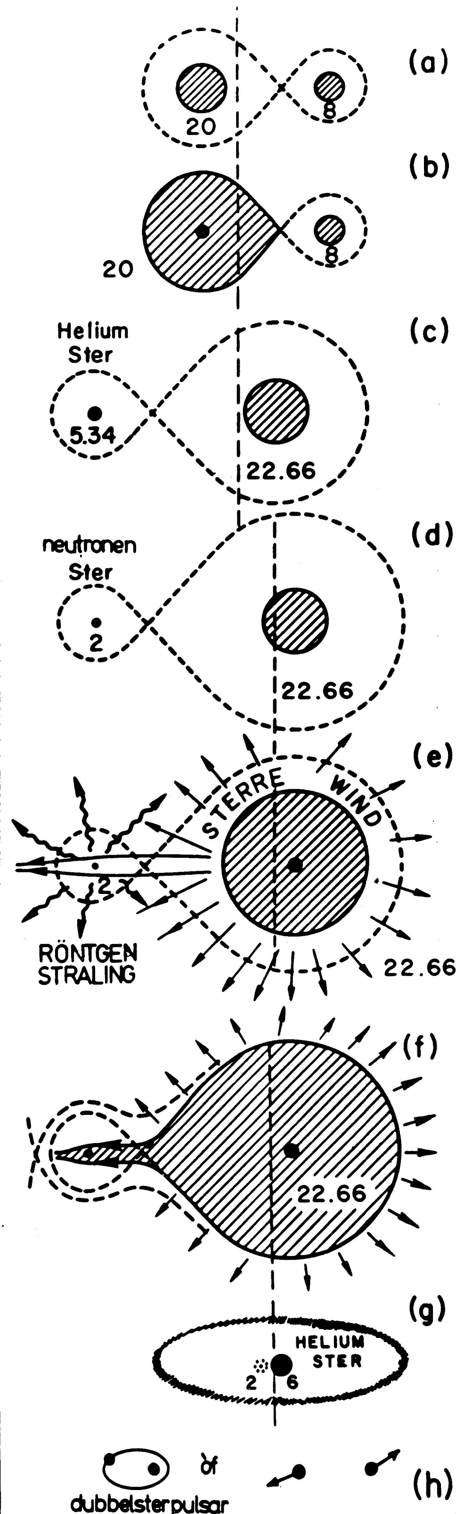
de gemeten betrouwbaarheidsgebieden van de vier parameters worden vastgelegd. Het eerste-orde-effect $\dot{\omega}$ is met zeer grote nauwkeurigheid gemeten en levert de teruglopende rechte op; in dit gebied is de onzekerheid kleiner dan de dikte van de lijn. Het gecombineerde gebied, opgeleverd door de drie hogere-orde-parameters is gearceerd. De grenzen van dit gebiedje zijn echter nogal onzeker. Men neemt dan ook aan dat deze twee gebieden niet significant verschillen. Alles samen levert: $M_p = 1{,}39 \pm 0{,}15\ M_\odot$ en $M_b = 1{,}44 \pm 0{,}15\ M_\odot$. ($M_\odot$ is de massa van de zon.) Merk op dat de pulsar en zijn begeleider vrijwel dezelfde massa hebben.

De aard van de begeleider

In bovenstaande analyse is verondersteld dat de begeleider zelf geen bijdrage levert tot de correcties op de keplerbaanparameters. Dit is alleen het geval indien deze begeleider zelf een *compacte* ster is, dat wil zeggen een neutronenster of een witte dwerg. Wanneer de begeleider een minder compacte ster is, zal door getijdewerking een zodanige vervorming van deze ster optreden dat het bijbehorende gravitationele quadrupoolmoment een aanzienlijke bijdrage tot $\dot{\omega}$ gaat geven. Op deze gronden kan men een gewone waterstofrijke hoofdreeksster (zoals de zon) direct uitsluiten omdat deze een periastrondraaiing van meer dan 3° per *omloop* zou leveren. Het quadrupoolmoment van een heliumster zou de gemeten $\dot{\omega}$ nog wel kunnen opbrengen. Echter, omdat in dat geval de relativistische bijdrage tot $\dot{\omega}$ slechts een kleine fractie daarvan kan zijn, zou een dergelijke heliumster slechts een geringe massa mogen hebben (slechts enkele tienden van een zonmassa). Dit levert dan weer problemen op met de gemeten waarde van $\dot{P}_b$, die meer dan zesmaal groter is dan men voor de zwaarst mogelijke heliumster kan verwachten. De enige mogelijkheden die voor de begeleider overblijven zijn dan: een neutronenster en (weliswaar minder waarschijnlijk) een zware, langzaam roterende witte dwerg. (Een snel roterende afgeplatte witte dwerg zou door zijn quadrupoolmoment een positieve $\dot{P}_b$ geven.)

Een recente optische identificatie[5] sluit een neutronenster uit. Hieraan wordt door de meeste astronomen geen geloof gehecht daar deze identificatie blijkt te berusten op een aanwijsbare, echter nog niet door iedereen begrepen fout in de gebruikte coördinatentransformatie.

Fig. 7. Mogelijke evolutieweg naar een dubbelsterpulsar[6]. In een dubbelster met massa's 20 en 8 $M_\odot$ zal de zwaarste ster het snelst evolueren (a) en massa overdragen aan de begeleider (b). De achtergebleven heliumster (c) explodeert als een supernova (d). Het restant is een neutronenster die eerst een röntgenpulsar zal worden (e) door de vangst van sterrewindmaterie en uiteindelijk als radiopulsar zal achterblijven (h). Intussen is de begeleider ook als heliumster aan zijn eind gekomen (g) en indien het systeem deze tweede supernova-explosie overleeft zal een dubbelsterpulsar kunnen overblijven (h), waarvan de jonge zowel als de oude neutronenster een radiopulsar kan zijn.

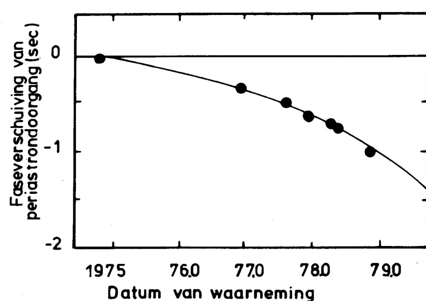


Er zijn ook evolutionaire overwegingen die een neutronenster als begeleider het meest aannemelijk maken (zie figuur 7). Berekeningen uitgevoerd in het Sterrenkundig Instituut van de Universiteit van Amsterdam tonen dat de voorloper van een dubbelsterpulsar een röntgendubbelster geweest zou kunnen zijn, bestaande uit een neutronenster die röntgenstraling uitzendt (de tegenwoordige radiopulsar) en een zware, zichtbare ster. Wanneer deze laatste ster aan het eind van zijn leven explodeert als een supernova zal er een tweede neutronenster achterblijven. Men verwacht dat in de meeste gevallen de dubbelster als gevolg van de supernova-explosie ontbonden zal worden. Maar àls de twee neutronensterren bijeen blijven wordt inderdaad een zeer excentrische baan verwacht. Tot nu toe is dit de enig mogelijke evolutieweg gebleken volgens welke een dubbelsterpulsar gevormd kan worden.

Gravitatiestraling

Taylor en zijn medewerkers voerden nog een test uit om de significantie van de afname van de baanperiode als gevolg van gravitatiestraling na te gaan. Hiertoe maakten ze de veronderstelling dat de baanperiode P_b in de loop van de tijd niet veranderd zou zijn. Onder deze veronderstelling losten ze wederom alle baanparameters op, waaronder ook het tijdstip van periastrondoorgang. Figuur 8 toont de aldus uit de waarnemingen afgeleide momenten van periastrondoorgang. De rechte lijn stelt het ware tijdstip voor van doorgang in 1974. De parabool stelt het teruglopen van de berekende momenten van doorgang voor indien de baanperiode afneemt als gevolg van gravitatiestraling volgens Einsteins theorie (indien beide sterren neutronensterren zijn met massa 1,41 $M_\odot$). De figuur toont duidelijk dat de afname van P_b in fraaie overeenstemming is met de algemene relativiteitstheorie.

Fig. 8. De periastrondoorgang als functie van de tijd. De punten geven het moment van periastrondoorgang zoals die geweest zou zijn als men bij de analyse de doorgang van 1974 (de rechte lijn) had gebruikt en de baanperiode constant had gehouden. De parabool komt overeen met de algemeen-relativistische voorspelde afname in de baanperiode als gevolg van gravitatiestraling.



Men bedenke dat deze gravitatiestraling *quadropool*straling is. Diverse gravitatietheorieën voorspellen ook het optreden van gravitationele *dipool*straling die (in het algemeen) sterker is dan de quadrupoolstraling. Helaas is in deze theorieën de intensiteit van de dipoolstraling precies nul indien de massa's van de twee sterren aan elkaar gelijk zijn, hetgeen hier niet uitgesloten is. Evenwel, de massabepalingen zijn verricht onder de vooronderstelling dat de algemene relativiteitstheorie (die *geen* dipoolstraling voorspelt) geldig is. Een analyse op grond van een andere theorie zal in het algemeen andere parameterwaarden opleveren, maar is nog niet gedaan. Het lijkt echter weinig waarschijnlijk dat zo'n analyse *precies* dezelfde waarden voor de massa's zal opleveren.

Kanttekeningen

Er is kritiek geuit op de juistheid van de formule voor gravitatiestraling die ten grondslag ligt aan de door Taylor *et al.* uitgevoerde analyse. In het bijzonder voert men aan dat de formule (Landau en Lifshitz geven de afleiding ervan in hun boek als opgave) niet exact is, doch slechts een benadering. De benadering ligt in het feit dat men in de berekening aanneemt dat de baan niet krimpt *tijdens* het uitzenden van de gravitatiestraling. Men brengt vervolgens dit energieverlies in rekening door de baan te laten krimpen. Dit is natuurlijk geen consistente aanpak van het probleem, maar tot nu toe ondervindt een exacte berekening zowel conceptuele als mathematische moeilijkheden. Toch geloven de meeste auteurs dat het slechts om een hogere-orde-correctie gaat, veel kleiner dan het voorspelde en waargenomen effect zelf. Hier ligt nog een uitdaging voor de mathematische fysici.

Verder kan worden geconstateerd dat in figuur 8 het eerste punt een nogal dominerend effect heeft op de kromming. Een paar jaar extra meten zal hier echter de twijfel kunnen wegnemen.

Op grond van bovenstaande kan men verwachten dat de dubbelsterpulsar hèt standaard testobject zal worden voor gravitatietheorieën. Tot nu toe is Einsteins algemene relativiteitstheorie in ieder geval in staat gebleken om alle gemeten niet-newtonse effecten kwalitatief èn kwantitatief te beschrijven. De komende jaren zullen leren welke andere gravitatietheorieën de toets aan de dubbelsterpulsar overleven.

De auteurs zijn drs. P. Hut zeer erkentelijk voor een verhelderende discussie over gravitatiestraling.

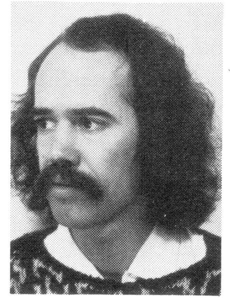
Referenties

1. R.A. Hulse, J.H. Taylor, Astrophysical Journal **195** (1975) L51.
2. J.H. Taylor, L.A. Fowler, P.M. Culloch, Nature **277** (1979) 437.
3. R.N. Manchester, J.H. Taylor: *Pulsars*. Freeman en Co. San Fransisco, 1977.
of ook: F.G. Smith: *Pulsars*. Cambridge University Press, Cambridge, England, 1977.
4. G.W.E. Beekman, Zenit **4** (1977) 2.
of ook: Sky and Telescope, maart 1975, 140.
5. P. Crane, J.E. Nelson, J.A. Tyson, Nature **280** (1979) 367.
6. E.P.J. van den Heuvel, in: *Physics and Astrophysics of neutron stars and black holes*. Redacteuren R. Giacconi en R. Ruffini. North Holland Publ. Co. Amsterdam, 1978.

Clusters van sterrenstelsels

N. Roos

Clusters van sterrenstelsels zijn de grootste objecten in ons expanderend heelal. Op een schaal groter dan 10 megaparsec (1 Mpc $\approx 3{,}1 \times 10^{22}$ m) kan de verdeling van materie als homogeen worden beschouwd. De snelheid waarmee sterrenstelsels (St's) met onderlinge afstand d (> 10 Mpc) van elkaar vandaan vliegen wordt dan gegeven door Hubble's wet: $v = H_0 d$ waarin H_0 de zgn. Hubble-constante is*. Op kleinere schaal blijken sterrenstelsels bij voorkeur in groepen of clusters van twee tot enkele honderden voor te komen. Het gemiddeld aantal heldere stelsels per Mpc^3 bedraagt ongeveer $2\text{x}10^{-3}$. Wanneer deze willekeurig in de ruimte waren verdeeld zou de gemiddelde afstand van een stelsel tot zijn naaste buur ongeveer 4,5 Mpc bedragen. In werkelijkheid is die afstand ongeveer 1 Mpc. De afstand van ons melkwegstelsel tot zijn naaste buur de Andromeda-nevel (M31) bijvoorbeeld is 0,7 Mpc; bovendien vallen ons melkwegstelsel en M31 naar elkaar toe met een snelheid van 50-100 km/s en niet van elkaar af zoals zou moeten in een uniform expanderend heelal. Samen met nog wat kleinere stelsels vormen ze een groep, de zgn. Lokale Groep van sterrenstelsels.**
In dit artikel zal het gaan om de vraag hoe groepen of clusters van sterrenstelsels ontstaan en hoe hun vormingsgeschiedenis samenhangt met die van sterrenstelsels.



Drs. Nico Roos (29) studeerde theoretische natuurkunde aan de VU te Amsterdam.
Sinds 1977 is hij verbonden aan de Rijksuniversiteit te Leiden (vakgroep sterrenkunde) waar hij bezig is met een promotieonderzoek aan clusters van sterrenstelsels.

Vorming van sterrenstelsels en clusters van St's

Sinds de ontdekking van de expansie van het heelal door Hubble, de ontdekking en interpretatie van de 3 K-achtergrondstraling, en de juiste voorspelling van de kosmische heliumabundantie staat het 'Big Bang' model stevig in zijn schoenen[1].
In dit kosmologisch model is het heelal geëvolueerd uit een zeer dichte, hete toestand, waarin straling en materie met elkaar in thermisch evenwicht waren. Door de expansie koelden straling en materie snel af. Het thermische evenwicht tussen hen werd verbroken bij $z \sim 1000$***, toen elektronen en protonen (re)combineerden. Men veronderstelt dat op dat moment fluctuaties met kleine amplitude aanwezig waren in de verdeling van materie. De expansie in zo'n dichtheidsverhoging wordt sterker afgeremd dan gemiddeld. De fluctuaties hebben daardoor de neiging te groeien (gravitatie-instabiliteit) en in sommige gebieden zal de expansie zelfs worden gestopt, waarna het gebied samentrekt totdat een evenwichtssituatie is bereikt. De mate van contractie wordt bepaald door de hoeveelheid energie die het gebied kwijt heeft kunnen raken. Zo kunnen massaconcentraties als sterren, sterrenstelsels en clusters van sterrenstelsels zijn ontstaan.
Het is mogelijk dat gaswolken met de afmetingen van grote clusters (M ongeveer 10^{13} à ^{14} $M_\odot$) zich als eerste losmaakten uit de kosmische expansie, en vervolgens instortten. De contractie hoeft zeker niet in 3 onafhankelijke richtingen tegelijkertijd te gebeuren. In de afgeplatte gebieden van hoge dichtheid ('pancakes') die zo ontstonden zouden de sterrenstelsels (St's) gevormd kunnen zijn[2].

In het alternatieve scenario gaat men ervan uit dat ook dichtheidsfluctuaties die minder massa omvatten kunnen groeien. Men neemt aan dat het spectrum van dichtheidsfluctuaties $\delta\rho/\rho$ na recombinatie voldeed aan

$$\delta\rho/\rho \propto M^{-\alpha}, \quad \alpha = 1/3 - 2/3.$$

In dit geval zullen de fluctuaties die de kleinste massa omvatten het eerst omkeren en instorten. Eerst worden de St's gevormd. Vervolgens ontstaan de groepen en clusters uit de fluctuaties in de verdeling van St's.
Een belangrijke vraag is die naar het onderscheid tussen St's en clusters van St's. Waarom komen er alleen St's voor met $M \lesssim 10^{12}$ $M_\odot$ en $R \lesssim 50$ kpc*. Het blijkt nu dat deze karakteristieke massa en straal direct volgen uit de eis dat een gaswolk door middel van vrij-vrij-interacties tussen protonen en elektronen zijn thermische energie zo snel kan uitzenden dat ongehinderde contractie onder invloed van zijn zwaartekracht kan plaatsvinden. De zo gevonden massa en straal van 10^{12} $M_\odot$ resp. 75 kpc zijn volledig uit te drukken in fundamentele fysische constanten[2]. Dissipatie van energie heeft zeer waarschijnlijk een belangrijke rol gespeeld bij de vorming van St's, terwijl de vorming van clusters alleen is bepaald door de zwaartekracht die de St's op elkaar uitoefenen.

Een belangrijk succes voor dit model is dat in numerieke simulaties van een expanderend heelal bestaande uit ongeveer 1000 massapunten de waargenomen clustering goed wordt gereproduceerd. Een belangrijke statistische maat voor de clustering van St's is de zgn. twee-puntscorrelatiefunctie $\xi(r)$. Deze functie kan als volgt worden gedefinieerd:

$$n(r) = \bar{n}\,\{1 + \xi(r)\},$$

waarbij $\bar{n}$ de gemiddelde dichtheid aan St's is en $n(r)$ de gemiddelde dichtheid op een afstand r van een St. Uit de waarnemingen volgt dat $\xi(r)$ over het hele gebied waarover zij is bepaald, 10 kpc $< r <$ 10 Mpc, goed wordt beschreven door een eenvoudige machtswet[3]:

$$\xi(r) = (r_0/r)^{1,8}, \quad r_0 \approx 10 \text{ Mpc}.$$

Het is dus niet mogelijk over een karakteristieke grootte van clusters te spreken.
Bij de kwantitatieve beschrijving van de clustering van St's vormt de zgn. 'multiplicity function' een goede aanvulling op de twee-puntscorrelatiefunctie. Deze verdelingsfunctie geeft het aantal clusters met totale optische helderheid L (hij neemt af met L volgens $L^{-2\pm0,3}$).
De correlatiefunctie die ontstaat in de numerieke simulatie van een expanderend heelal heeft dezelfde vorm als de waargenomen functie, en ook de waargenomen en berekende multiplicity-functies komen

* We zullen in dit artikel aannemen dat H_0 = 50 km·s^{-1}. Mpc^{-1}.
** Onder 'heldere' sterrenstelsels zullen we verstaan: St's met optische helderheid $L \sim 3\text{x}10^{10}$ $L_\odot$
$L_\odot$ is de helderheid van de zon ($\approx 3{,}86\text{x}10^{26}$ watt). De massa van de zon $M_\odot \approx 2 \times 10^{30}$ kg.
*** In Big Bang-kosmologieën hangt de roodverschuivingsparameter $z = \Delta\lambda/\lambda$ ($\Delta\lambda = \lambda_0 - \lambda$, λ is de golflengte van een foton op het moment van uitzending en λ_0 de golflengte bij ontvangst) direct samen met de schaal R van het heelal volgens $R/R_0 = (1 + z)^{-1}$.
Het verband tussen R en de leeftijd van het heelal t hangt af van de dichtheid in het heelal, maar wordt (in ieder geval in het vroege heelal) goed beschreven door $R/R_0 = (t/t_0)^{2/3}$. De huidige leeftijd van het heelal, t_0, wordt geschat op 10 à 20 miljard jaar.

* Uitzonderingen op deze regel vormen zgn. cD (central Dominant)-stelsels, welke waarschijnlijk een bijzondere vormingsgeschiedenis achter de rug hebben (zie volgende paragraaf). De massa van deze stelsels is ongeveer 10^{13} $M_\odot$.

goed overeen[4]. De mate van clustering aan het eind van de simulatie, bepaald door r_0, hangt af van de waarde van z waarop de berekening wordt gestart, ofwel de roodverschuiving waarop St's worden gevormd. Dat blijkt ongeveer bij z = 20 te moeten gebeuren.
Begincondities zoals de vorm van het fluctuatiespectrum en de dichtheid van het heelal blijken weliswaar geen sterke, maar toch meetbare invloed te hebben op de eindsituatie, en het ziet ernaar uit dat verdere studie op dit gebied veel kan verduidelijken over de omstandigheden in het vroege heelal en de ontstaansgeschiedenis van St's en clusters van St's.

De rijke clusters

Classificatie
Abell heeft in 1958 een catalogus samengesteld van 2712 rijke clusters van sterrenstelsels, geïdentificeerd op platen van de Palomar Sky Survey. Deze clusters hebben roodverschuiving z = 0,02 - 0,2 en bevatten 20 tot 100 heldere leden binnen een afstand van 3 Mpc tot het clustercentrum. Heldere sterrenstelsels kunnen worden onderscheiden in 3 typen: Spiraalstelsels (Sp), SO-stelsels, welke beide bestaan uit een platte roterende schijf met een rond of ellipsvormig lichaam (de 'bulge') in het centrum, en elliptische stelsels (E). SO-stelsels bevatten evenals elliptische stelsels zo goed als geen (neutraal) gas en hebben over het algemeen wat grotere bulges dan spiraalstelsels. Er bestaat een verband tussen de structuur van rijke clusters en de fractie spiraalstelsels die ze bevatten[5]. Sommige clusters bevatten overwegend SP-stelsels (ongeveer 60%), 25% SO-stelsels en 15% E-stelsels (deze percentages vindt men ook in de gebieden buiten de rijke clusters) De structuur van deze clusters is meestal onregelmatig en er zijn vaak verschillende massaconcentraties in te onderscheiden. Hun tegenhangers hebben meestal een regelmatige, bijna sferisch-symmetrische vorm met een goed gedefinieerd centrum waarin de dichtheid aan sterrenstelsels heel hoog kan zijn (1000 Mpc^{-3}). Deze clusters bestaan overwegend uit SO- en E-stelsels en bevatten weinig Sp-stelsels. In alle clusters neemt de fractie elliptische stelsels toe met de dichtheid aan sterrenstelsels.
In de buurt van het centrum van veel regelmatige clusters met een hoog percentage E-stelsels, vindt men vaak een enorm groot en helder elliptisch stelsel: een zogenaamd cD (central Dominant)-stelsel. De massa van deze stelsels kan gemakkelijk $10^{13}\ M_\odot$ bedragen.

Fig. 1. Een deel van de Hercules-cluster. Dit is een typisch voorbeeld van een (open) onregelmatige, spiraalrijke cluster. Opvallend veel leden zijn vervormd door wisselwerking met een ander stelsel. De foto is 33' breed. Op de afstand van Hercules (~ 200 Mpc) komt dat overeen met ~ 2 Mpc.



Instorting
In numerieke simulaties van een expanderend heelal, zoals in de vorige paragraaf genoemd, manifesteert clustering zich het eerst op kleine en vervolgens op steeds grotere schaal. Simulaties van de evolutie van rijke clusters laten ook sterke subclustering zien tijdens de expansie en de daarop volgende instorting van de cluster. Gedurende deze periode, welke ongeveer $t_c \approx (G\rho)^{-1/2}$ duurt (G = gravitatieconstante, ρ = dichtheid van de cluster) krijgen de sterrenstelsels steeds grotere 'random' snelheden door wisselwerking met het voortdurend veranderend gravitatieveld. Dit proces staat bekend als violent relaxation. Al na ongeveer $3/2\ t_c$ is er een evenwicht tussen kinetische (T) en potentiële energie (V) van de cluster ontstaan, beschreven door het viriaaltheorema: $2T + V = 0$. De subclustering is dan uitgewist en de cluster heeft een regelmatig dichtheidsverloop. Dit suggereert dat twee extreme clustertypen uit de vorige paragraaf verschillende evolutiestadia representeren.

Missing mass
Uit metingen van de snelheden van een aantal sterrenstelsels in een cluster kan met behulp van het viriaaltheorema worden de clustermassa worden geschat. Deze blijkt dan ongeveer een factor 10 groter te zijn dan de som van de massa's van de sterrenstelsels. Alleen de massa in het optisch zichtbare deel van de stelsels wordt dan meegerekend. Metingen van de rotatiesnelheid van neutraal waterstof buiten het zichtbare deel van de schijf van spiraalstelsels wijzen erop dat deze stelsels zijn omringd door donkere zware halo's. Deze halo's zouden kunnen bestaan uit objecten die al in een vroeg stadium van de vorming van de stelsels ontstonden. De massa M_g van de stelsels blijkt evenredig met de straal R_g toe te nemen. Wanneer nu de 'missing mass' van de clusters aan de sterrenstelsels vastzit, zijn deze stelsels ongeveer 10 maal zo groot en zwaar als men aanvankelijk dacht.

Botsingen, en de vorming van elliptische stelsels
Zeker wanneer St's zijn omringd door massieve halo's zullen botsingen tussen St's veel voorkomen in rijke clusters. De gemiddelde tijd t_{coll} die verloopt tussen 2 opeenvolgende botsingen van een stelsel kan worden geschat met $t_{\mathrm{coll}}^{-1} \sim \pi R_g^2\, n_{\mathrm{cl}}\, \sigma_{\mathrm{cl}}$ waarin n_{cl} de dichtheid aan St's en σ_{cl} de snelheidsdispersie van de stelsels in de cluster is. In het centrale deel van een dichte, regelmatige cluster ($n_{\mathrm{cl}} \sim 10^{-3}\ \mathrm{Mpc}^{-3}$, $\sigma_{\mathrm{cl}} \approx 10^3$ km/s) is dat typisch 10^8 tot 10^9 jaar voor een stelsel met $R_g \sim 10$ kpc. Voor grotere stelsels (halo's) is die tijd aanzienlijk korter. Deze botsingstijd moet natuurlijk worden vergeleken met de leeftijd van zo'n cluster welke $\sim 5 \times 10^9$ jaar bedraagt.

Tijdens een ontmoeting tussen 2 sterrenstelsels met relatieve snelheid op grote afstand v en afstand van dichtste nadering p, wordt op de sterren een getijdekracht uitgeoefend, welke de interne energie U van het stelsel verhoogt. Wanneer $v \gg \sigma_g$, de interne snelheidsdispersie van het stelsel, en $p > R_g$ volgt met behulp van de aanname dat de sterren zich tijdens de ontmoeting niet ten opzichte van het centrum van het stelsel verplaatsen (de zgn. impulsbenadering) dat $\Delta U/U \propto (R_g/p)^4 (\sigma_g/v)^2$. Botsingen met lage v en p hebben dus het grootste effect. Bij lage snelheden is de impulsbenadering niet betrouwbaar meer. Er treden dan gecompliceerde vervormingen van het stelsel op tijdens de botsing welke alleen numeriek goed gesimuleerd kunnen worden. De toename van de interne energie gaat ten koste van de baanenergie en de botsingen worden sterk inelastisch bij lage snelheden. Uit numerieke berekeningen blijkt dat twee identieke stelsels, die botsen met $p < R_g$ en $v < \sigma_g$, elkaar binnen korte tijd na de eerste ontmoeting invangen en een nieuw stelsel vormen met $\sigma \approx \sigma_g$.

De inelasticiteit van sterrenstelsels kan gemakkelijk op een schematische manier worden ingebouwd in numerieke simulaties. Dergelijke berekeningen tonen aan dat tijdens de vorming van clusters, wanneer σ_{cl} nog laag is, een aanzienlijke fractie van de sterrenstelsels naar zijn meest nabije buur toevalt en wordt ingevangen. In Arp's 'Atlas of peculiar galaxies', komen veel dubbelstelsels voor die blijkens hun sterk afwijkende uiterlijk grote getijdekrachten op elkaar uitoefenen. In 1972 werd voor het eerst door de gebroeders Toomre gesuggereerd dat deze stelsels elkaar binnen zo'n 5×10^8 jaar zouden invangen en dan gezamenlijk een *elliptisch* stelsel zouden vormen. De numerieke experimenten hebben het inmiddels op verscheidene gronden aannemelijk gemaakt dat *alle* elliptische stelsels (behalve misschien de zgn. dwerg-elliptische stelsels) het resultaat zijn van invangst[6]). De sferische component (SC, de 'bulge') van spiraalstelsels en SO's vertonen veel overeenkomsten met elliptische stelsels. Recente waarnemingen hebben aangetoond dat de grootte van de SC gecorreleerd is met de morfologische klasse waartoe een St behoort (SO-stelsels hebben grotere bulges dan spiraalstelsels). Het is mogelijk dat alle heldere St's zijn gevormd als schijfvormige (spiraal-)stelsels met nauwelijks of geen SC en dat vervolgens, door middel van invangst van kleine stelsels, de SC's en daarmee de huidige typen St's ontstonden. Het feit dat invangst (en waarschijnlijk het ontstaan van SO- en E-stelsels) vooral plaatsvindt tijdens het ontstaan ($t < t_c$) van clusters vormt dan weer een aanwijzing dat de hierboven genoemde clustertypen verschillende evolutiestadia representeren.
De cD-stelsels in de centra van regelmatige, spiraal-arme clusters lijken hun enorme afmetingen te danken te hebben aan het feit dat ze een stuk of tien St's hebben opgeslokt. Door het grote aantal botsingen zijn de meeste stelsels in clusters het grootste deel van hun halo's kwijt. Ze bewegen nu door een zee van donker halomateriaal. Door hun wisselwerking met deze materie verliezen vooral de grote stelsels voortdurend baanenergie. Dit proces staat bekend onder de naam 'dynamical friction' en is het gevolg van de tendens equipartie van energie te bewerkstelligen tussen de stelsels en het lichtere achtergrondmateriaal. Zo zullen de zware stelsels in een rijke cluster de neiging hebben naar de bodem van de potentiaalput te spiraleren om daar opgegeten te worden door het grote centrale cD-stelsel.

Intergalactisch gas
De meerderheid van geïdentificeerde extragalactische röntgenbronnen zijn geassocieerd met rijke clusters van sterrenstelsels. Het belangrijkste emissiemechanisme is remstraling afkomstig van geïoniseerd gas met een temperatuur van $T \sim 10^7$ à ^{8} K en dichtheid $\sim 10^{-3}$ protonen/cm^3.

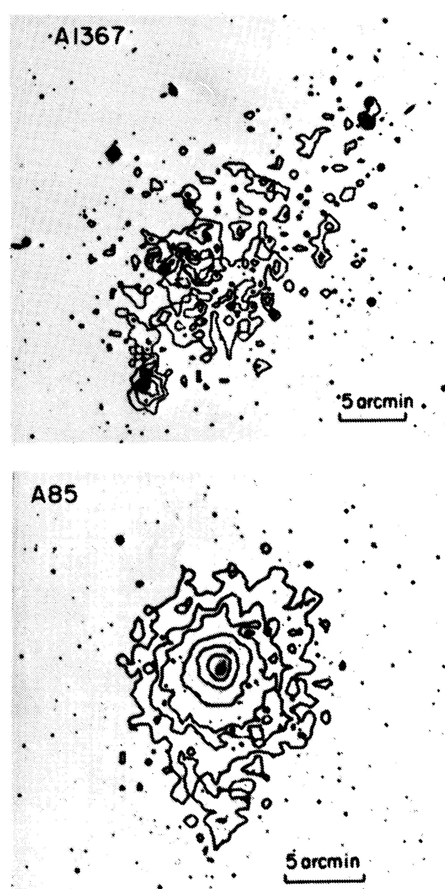


Fig. 2. Twee rijke clusters (in negatief afgedrukt) met hun röntgenisofoten. De röntgenwaarnemingen zijn onlangs gedaan met de HEAO-2 (High Energy Astrophysical Observatory) satelliet. A1367 is een onregelmatige, spiraalrijke cluster, terwijl A85 een centraal geconcentreerde cluster is met een hoog percentage aan elliptische stelsels. Ook de röntgenwaarnemingen laten een duidelijk verschil in structuur tussen de twee clusters zien.

Deze temperatuur is ongeveer gelijk aan de viriaaltemperatuur: $kT_{vir} = GM_{cl}/R_{cl}$ (M_{cl} en R_{cl} zijn de massa en straal van de cluster). De totale massa van het gas is $\sim 0{,}1\ M_{cl}$. Het bestaan van dit gas blijkt ook uit radiowaarnemingen van sterrenstelsels in clusters. Radiostelsels zenden in twee tegenoverliggende richtingen wolken uit welke een mengsel van relativistische deeltjes en magneetvelden bevatten en daardoor zichtbaar worden in het radiogebied (synchrotronstraling!). Wanneer deze stelsels zich met hoge snelheid door het intergalactisch gas bewegen worden deze wolken naar achteren gedrukt ('ram pressure') en we zien een zgn. 'kop-staart-stelsel' of 'radio-spoor'. De druk die het gas uitoefent langs de bewegingsrichting van de sterrenstelsels kan het gas uit spiraal stelsels 'wegvegen'. Zo'n schoongeveegd stelsel zal als een SO-stelsel worden geclassificeerd.
Er zijn aanwijzingen dat dit inderdaad gebeurt in de centra van rijke clusters. Onlangs is de aanwezigheid van een emissielijn afkomstig van 15x geïoniseerd Fe ontdekt in het röntgenspectrum van het clustergas. De Fe-abundantie die hieruit is af te leiden geeft aan dat het gas grotendeels van stellaire herkomst is (elementen zwaarder dan He zijn in sterren geproduceerd, niet in de Big Bang). Het is goed mogelijk dat de halo's van St's hierbij een belangrijke rol hebben gespeeld. Het galactische (halo)gas kan door de processen die samenhangen met de evolutie van de cluster, zoals instorting van de cluster, botsingen en ram pressure, uit de stelsels zijn vrijgemaakt en verhit tot zijn huidige temperatuur.

Waarnemingen nu en in de nabije toekomst
Het in november 1978 gelanceerde Einstein-observatorium of HEAO-2 (High Energy Astrophysical Observatory) is in staat röntgenwaarnemingen te doen met een hoog ruimtelijk en spectraal scheidend vermogen (zie figuur 2). De verwachting is dat deze röntgentelescoop individuele clusters zal kunnen onderscheiden tot aan een roodverschuiving van $z \approx (1-3)$. De evolutie van het clustergas kan hiermee direct worden waargenomen. Dit is van belang voor het antwoord op de vraag welke het eerst ontstonden: sterrenstelsels of clusters van sterrenstelsels, en hóe ze ontstonden.

Recente optische waarnemingen van verre clusters ($z \approx 1$) laten zien dat een groot aantal sterrenstelsels in die clusters relatief erg blauw zijn. In de centrale delen van nabije, rijke clusters komen overwegend E-stelsels voor, welke een rodere kleur hebben. De kleur van de stelsels in de clusters bij $z \sim 1$ is veel meer karakteristiek voor spiraalstelsels. Ook radiostelsels lijken een sterke kleurevolutie te ondergaan. Het ligt voor de hand dat er een verband is tussen de evolutie van sterrenstelsels en de sterke z-afhankelijkheid die wordt waargenomen in het aantal radiobronnen en quasars in het heelal.

In 1983 wordt de Space Telescope gelanceerd. Met dit instrument zullen sterrenstelsels tot aan $z \sim 1$ goed bestudeerd kunnen worden over een breed gedeelte van hun spectrum (UV tot aan het nabije IR). Het ziet ernaar uit dat we binnen afzienbare tijd de vorming van sterrenstelsels en clusters van sterrenstelsels direct kunnen waarnemen.

Referenties

1. S. Weinberg: *Gravitation and Cosmology*. John Wiley and Sons, New York, 1972.
2. M.J. Rees, in *Observational Cosmology*, 8th Advances Course, Saas Fee, 1978. Publ. Geneva Observatory, CH-1290 Sauverny, Switzerland.
3. P.J.E. Peebles, Astrophys. Journal **189**, (1974) L51.
4. J.R. Gott, E.L. Turner, S.J. Aarseth, Astrophys. Journal **234**, (1979) 13.
5. N.A. Bahcall, Ann. Rev. of Astron. & Astroph. **15** (1977) 505.
6. N. Roos, C.A. Norman, Astron & Astroph. **76** (1979) 75.